교육학석사학위논문

The Effects of Question/Option Presentation Mode and Item Type on L2 Listening Test Performance and Perception of Korean EFL Leaners

문항 제시 방법과 문항 유형이
한국인 영어 학습자들의 청해 시험 수행과
그에 대한 인식에 미치는 영향

2016년 8월

서울대학교 대학원

외국어교육과 영어전공

염 수 혜

The Effects of Question/Option Presentation Mode and Item Type on L2 Listening Test Performance and Perception of Korean EFL Leaners

by
SOOHYE YEOM

A Thesis Submitted to
the Department of Foreign Language Education
in Partial Fulfillment of the Requirements
for the Degree of Master of Arts in Education

At the
Graduate School of Seoul National University
August 2016

The Effects of Question/Option Presentation Mode and Item Type on L2 Listening Test Performance and Perception of Korean EFL Leaners

문항 제시 방법과 문항 유형이
한국인 영어 학습자들의 청해 시험 수행과
그에 대한 인식에 미치는 영향

지도교수 소 영 순

이 논문을 교육학 석사 학위논문으로 제출함
2016년 5월

서울대학교 대학원
외국어교육과 영어전공
염 수 혜

염수혜의 석사학위논문을 인준함
2016년 8월

위 원 장 _____
부위원장 _____
위 원 _____

The Effects of Question/Option Presentation Mode and Item Type on L2 Listening Test Performance and Perception of Korean EFL Leaners

APPROVED BY THESIS COMMITTEE:

_____

BYUNGMIN LEE, COMMITTEE CHAIR

_____

YOUNGJU LEE

_____

YOUNGSOON SO

Abstract

This study explored the effects of question/option presentation mode and item type on EFL learners' listening comprehension performance and their perception. The aural and written modes of presentation are compared for the two item types, the dialogue-completion and question-and-answer (Q&A). One hundred and fifteen Korean college students participated in the study, and they were divided into three different proficiency groups: low intermediate, mid/high intermediate, and advanced levels. The participants took a listening test with 4 aural and 4 written dialogue-completion multiple-choice items and another 4 aural and 4 written Q&A multiple-choice items. After taking the test, the participants also completed a survey on their perceptions of each section of the test and participated in a stimulated recall interview.

The results showed that the low intermediate group performed significantly better in the written mode than in the aural mode, while they received similar scores on the two item types. This coincides with the survey results, in which the aural mode was perceived as more difficult, and the written mode was preferred. Moreover, unlike the groups with higher proficiency levels who regarded the aural mode more valid for both the

dialogue-completion and Q&A items, the low intermediate group felt the aural mode was better for the dialogue-completion items, while the written mode was more appropriate for the Q&A items. The survey and the stimulated recall interviews revealed that the reason the lowest proficiency group found the aural mode much more challenging was that it, additionally, required a high level of concentration and good memory. This did not mean that the written mode was better for them, however, because it allowed them to use the word-matching strategy which was not relevant to the listening comprehension. Also, reading the options in the allotted amount of time was difficult for many low proficiency participants when taking the test in the written mode.

On the other hand, while the mid/high intermediate and advanced group scored significantly higher on the dialogue-completion items than on the Q&A items, they demonstrated little difference in scores between the two modes. Still, they felt the aural mode was much more difficult than the written mode, and the advanced group even expressed a stronger preference for the written mode than the other two groups did. The results from the survey and the stimulated recall interview suggested that even though they could perform equally well on the both modes, they were more confident about reading the questions and options than listening them. Also, they did not have any

difficulty in reading quickly, which allowed them to read the options in advance and predict the listening stimuli.

All things considered, memory capacity and reading ability were found to be the two major construct-irrelevant factors in the aural and written mode of the listening comprehension tests, respectively. Since we do not want the memory capacity or the reading ability to decide the listening test results in most cases, the effort to minimize their effects is necessary when developing the tests. The implications and limitations of this study and suggestions for future research are discussed.

Keywords: multiple-choice questions, second language listening, test format

Student Number: 2014-22961

TABLE OF CONTENTS

List of Tables

Chapter 1.

INTRODUCTION

This chapter introduces the research by presenting the motivation of the study and the organization of the thesis. Section 1.1 discusses the background and the purpose of the study. Section 1.2 presents the research questions, and the overall organization of the thesis is outlined in Section 1.3.

1.1. Background and Purpose of the Study

The extent to which different testing methods affect the performance of test-takers has been an important issue in developing language tests and interpreting the scores obtained from the tests (Bachman, 1990). This is mainly due to the strong influence of the test method on the test validity. Since tests seek to measure specific constructs, the degree to which a test can be considered valid depends on how aptly it can measure a specific construct. However, different measurement characteristics can invite different test-taking processes and constructs that are measured through the test can be changed, which thus can yield varying outcomes.

Investigation of the effect of test methods is crucial for testing listening comprehension, since the test-takers' mental processing on the receptive skills (listening and reading) is not observable. When assessing listening, test-takers do not produce language and therefore it cannot be directly analyzed, so inferences about the test-takers' listening comprehension abilities are drawn from the scores that are obtained from certain test methods. Moreover, testing second language (L2) listening is even more complicated, inviting various factors that affect the test-takers' comprehension and performance. For example, processing a second language requires better memory of listeners than processing their first language (Cook, 2013; Ohata, 2006). In this case, a L2 listening test may measure test-takers' memory capacity in addition to their listening competence.

One of the most frequently used testing methods for assessing listening is multiple-choice question (MCQ) items, which can take different forms, depending on specific features, such as the ways of presenting instructions, questions, and options. For example, the options can be provided aurally or in written form. Although several previous studies revealed what factors may affect the difficulty of listening comprehension tests (Brindley & Slatyer, 2002; Buck, 1990, 2001; Freedle & Kostin, 1996; Nissan, DeVincenzi, & Tang, 1995), not many have focused on each characteristic

and the extent to which it impacts the test-takers performance. To gain a clearer and more detailed picture of test method effect on MCQ tests, the present study focuses on the effect of the question/option presentation modes and different MCQ item types among various factors that affect test-takers' listening MCQ test performance. The questions/option presentation modes are written or aural, and the item types of focus in this study were dialogue-completion and question-and-answer (Q&A). A more detailed explanation about these features will be provided in Chapter 2.

Making decisions on whether to use aural or written mode of question presentation is an important issue when developing a listening test, in that it is directly related to construct validity. Bachman (1990) emphasized that "in examining the effects of test method facets on language test scores, we are also testing hypotheses that are relevant to construct validity" (p. 258). The question/option presentation mode is one facet of the test method, and depending on the mode, other factors which are not directly relevant to the constructs that are aimed to be measured can influence the results. Therefore, to measure what is intended to be measured in the listening test, the presentation mode needs to be considered in development process (Bachman & Palmer, 1996). That is, construct-irrelevant method variance (Messick, 1996) should be minimized for the listening test to actually measure the test-

takers' listening ability.

Construct-irrelevant factors found in the written mode of question and option presentation include reading ability, lexical attractiveness (Freedle & Fellbaum, 1987; Freedle & Kostin, 1996, 1999) and uninformed guessing (Wu, 1998). Test-takers' ability to listen and read at the same time can also affect the result in this mode. Chang and Read (2013) and Yanagawa and Green (2008) raised the possibility of the written mode negatively affecting the validity of the listening test, based on their results that test-takers with lower listening proficiency performed significantly better with the written form. They concluded that if lower level test-takers did well because of the written questions and options, a construct-irrelevant factor, the reading ability, might have been measured in the written mode. This could make the written mode less valid for a listening test. Aural mode, on the other hand, requires short-term memory capacity, which is known to be particularly limited for L2 learners (Cook, 2013; Ohata, 2006). It is also found that the aural mode increases test-takers' anxiety which could affect their performance (Buck, 1991; Chang & Read, 2006).

Modes of question/option presentation also affects authenticity of the test and test-takers' cognitive processes. Bachman (1990) stated that

"approaches to authenticity are concerned with the context and manner in which we elicit a sample of performance – with the characteristics of the testing methods we use" (p. 303). That is, when determining each characteristics of the test method, how it reflects the real-life or authentic listening situations should be carefully examined, considering the target language use domain and the purpose of the test (Bachman & Palmer, 1996). Providing only answer options in written form, for example, has been criticized in previous studies (Hemmati & Ghaderi, 2014; Yanagawa & Green, 2008), since it exposes listeners to options before knowing what to listen for or listening to the stimuli, which is far from what we do in the real-life listening situation.

For developing more authentic MCQ listening tests that actually assess listening competence, it is crucial for test developers to make well-grounded decisions on the presentation mode. The current study, therefore, seeks to investigate the effect of the two different modes of question/option presentation on test-takers' L2 listening performance with two item types.

To investigate the influence of item type, the two most widely-used MCQ item types were chosen to be examined for the present study: dialogue-completion and question-and-answer (Q&A) item type. The item type is one

of the test method characteristics, and it has been considered to have effect on the test-takers' test performance (Berne, 1992; Wolf, 1993). Although both the dialogue-completion items and the Q&A items are MCQ items and both aim to measure the test-takers' listening comprehension, the two item types assess their listening ability in different ways. Test-takers complete a dialogue for a dialogue-completion item, while they answer a question about a dialogue for a Q&A item. The two types involve different listening skills and this can differently affect the test-takers' performance. Moreover, there is a possibility of dissimilar effects when these two types are combined with the two question/option presentation modes, aural or written.

In addition, proficiency levels of test-takers are taken into considerations to see if different proficiency groups respond to the mode and item type differently. Previous studies indicated that the test performance was differently affected by the test method characteristics depending on the test-takers' proficiency levels, and mixed results were found particularly for the lower-level test-takers (e.g. Chang & Read, 2013; Yanagawa & Green, 2008). Therefore, in this study, how the lower-level test-takers react to different presentation modes and item types are closely examined compared to the performance of higher-level test-takers.

Summarizing the purpose of the present study, research questions are stated in the following section.

1.2. Research Questions

The present study examines how differing test method characteristics of MCQ tests affect L2 listening test-takers' performance, depending on two independent variables: modes of presentation and item types. Since previous research found mixed results regarding the effect of aural and written modes of question presentation, particularly for lower-level test-takers, the listening test scores of three groups with different proficiency levels are taken into account. Test-takers' perceptions are also investigated using a survey and an in-depth verbal report. The following three research questions are addressed in this study:

1) To what extent do modes of question/option presentation and item type affect three proficiency groups' L2 listening performance?

2) What is the three proficiency groups' perception of aural and written modes of presentation on two item types?

1.3. Organization of the Thesis

The present study consists of six chapters. Chapter 1 introduces the purpose of the study and presents the research questions. Chapter 2 presents the literature review on effects of modes of question/option presentation and item types. In Chapter 3, the method of the study is described in terms of the participants, the instruments, the procedure, and the data analysis. Chapter 4 presents the results and the research findings are discussed in Chapter 5. Finally, Chapter 6 concludes the research with the summary of the major findings, the implications of the study, and the suggestions for future research.

Chapter 2.

LITERATURE REVIEW

Identifying various factors that may affect listening item difficulty has been the central interest in a number of studies (Buck, 1990, 2001; Freedle & Kostin, 1996; Nissan et al., 1995). Based on previous research, Brindley and Slatyer (2002) grouped the factors affecting task difficulty of listening assessment into three categories: the nature of the input, the nature of the assessment task, and individual listener factors. Each category includes factors listed below.

- **The nature of the input:** speech rate, length of passage, syntactic complexity, vocabulary, discourse structure, noise level, accent, register, propositional density, amount of redundancy, etc.
- **The nature of the assessment task:** amount of context provided, clarity of instructions, response format, availability of question preview, etc.
- **The individual listener factors:** memory, interest, background knowledge, motivation, etc. (Brindley & Slatyer, 2002, p. 375)

Among the three categories, factors related to the nature of the assessment task have been of great importance in language testing research. The testing instrument used for testing language is the language itself that is assessed (Bachman, 1990), and this complexity of testing language has led many researchers to investigate the extent to which different testing methods that aim to measure the same trait affect the performance, or the scores, of the test-takers. While some studies focused on different item types such as multiple-choice, open-ended, or cloze questions (e.g. Berne, 1992; Cheng, 2004; Hansen & Jensen, 1994), others investigated the differing effect of other item characteristics, including presentation mode, question preview, response format (e.g. Berne, 1995; Chang, 2008; Iimura, 2010). The two factors, the question/option presentation mode and the item type, that the present study is focusing on are included in this category as well, since they are test method characteristics.

In this section, studies on these two particular kinds of item characteristic variables will be reviewed: presentation modes and item types. Table 2.1 provides some examples to illustrate the two different modes of question/option presentation (aural and written mode) and the two item types (dialogue-completion and Q&A). The dialogue-completion and the Q&A item type items are presented in either aural or written mode.

Table 2.1

Sample Items from Two Item Types in Aural and Written Mode

Item Type	Listening Stimuli (Presented aurally)	Question/option Presentation Mode (How it is presented on test paper)	
		Aural Mode	Written Mode
Dialogue-	1. M: How do I get to the library? W: Follow this road to the big brick building. M: And that's the library? W: _____	1. (a) (b) (c) (d)	1. Choose the most appropriate response to complete the conversation. (a) It's my book. (b)No, but I'll try. (c) Yes, that's it. (d) Study them first.
Q&A	2. M: Are you transferring to a different university? W: Yeah, to one that's closer to home. M: Oh? Which one? W: Scarborough University. It's located near my family and has my major. M: That sounds good. W: Yes, I'm happy about it.	2. (a) (b) (c) (d)	2. What is the main topic of the conversation? (a) A plan to change schools (b) The high cost of education. (c) A comparison of two universities (d) Programs at Scarborough University

In both aural and written modes, listening stimuli are presented aurally. Questions and options, however, are presented differently in the two modes, as shown in the Test Paper section in Table 2.1. For the aural mode, both the question and the options are given aurally only, and there is nothing on the test paper other than (a), (b), (c), and (d) for marking the answer. For the written mode, on the other hand, both the question and the options are written on the paper. For a dialogue-completion item, test-takers listen to a short conversation and choose the most appropriate response to complete the conversation. Q&A items are different from dialogue-completion items in that the test-takers choose the correct answer for a question about the conversation.

Mode of presentation has been one of the controversial issue in developing listening comprehension tests, because written multiple-choice listening test items require reading questions and options which is irrelevant to the construct that the listening test is measuring. Also, item type is another characteristic of interest, since few studies considered it when examining the impact of two different modes of presentation. Previous research on these two item characteristics, presentation mode and item type, are discussed in Section 2.1 and 2.2, respectively, and the potential interaction between the two characteristics is reviewed in Section 2.3.

2.1. Modes of Question and Option Presentation

Before discussing previous research on the modes of question/option presentation in listening tests, some key terms need to be clarified, since different terms were used in the previous studies to explain the aural and the written mode of presentation. The work of Hemmati and Ghaderi (2014) and Yanagawa and Green (2008) focused on previewing effect of the written form of the questions and options, referring to the aural mode as 'non-preview', and the written mode as 'preview'. Chang and Read's (2013) study used different terms, contrasting the delivery mode of the questions and options. In this case they chose to use the terms 'oral mode' and 'written mode.'

Since the focus of the present study is not restricted to the previewing effect of the written mode, it employed the terms 'aural mode' and 'written mode' from the test-takers' perspective, whether they listen to the questions and options or read them. In fact, previewing is not an exclusive property of the written mode. It is also possible in the aural mode to some extent, in that the question can be played before the listening stimuli to have previewing effect. Therefore, the present study focused on whether the test-takers listen or read the question and option, not the previewing effect itself. For an effective comparison of the presentation modes, questions are played once

more before the listening stimuli in the aural mode to have a similar previewing effect to the written mode. A more detailed explanation about the development of the listening material will be provided in Section 3.2.1.

Although previous research generally agreed that providing questions or options in written form has a great effect on test-takers' performance (e.g. Chang, 2008; Iimura, 2010; Wu, 1998; Yanagawa & Green, 2008), they have yielded mixed results on the direction of its influence. Studies that showed positive effects of written questions or options explained that previewing questions in the written mode, which was not possible in their aural mode items, was beneficial for test-takers, since it motivated them by providing contextual information and relevant clues of the listening input and letting them employ metacognitive strategies such as goal setting and planning (Berne, 1995; Buck, 1991; Iimura, 2010; Yanagawa & Green, 2008).

Ur (1984) and Weir (1993), however, claimed that written questions and options distract test-takers' attention on listening input, since providing item stems and options in written form requires test-takers' reading ability. It was also discovered that written answer options invites uninformed guessing (Wu, 1998) and increases lexical attractiveness (Freedle & Kostin, 1996, 1999; Yanagawa & Green, 2008), letting listeners fall back on a lexical matching

strategy. In other words, when the test-takers were provided with answer options in written mode, they tended to just match words from the listening stimuli to the written options, without understanding their meaning.

In addition to the lexical matching strategy, Yanagawa and Green (2008) and Hemmati and Ghaderi (2014) also pointed out that the written answer options "provides contradictory cues and complicating planning strategies" (Yanagawa & Green, 2008, p. 110). This point is closely related to the authenticity of the test, which implies that the test-takers' cognitive processes should also be comparable to the cognitive processes that listeners employ in non-assessment situations. Since test-takers' cognitive processes in listening tests is influenced by the characteristics of test methods (Bachman & Palmer, 1996; Weir, 2005), providing the written answer options can undermine the validity of a listening test, as it may not reflect an authentic listening process.

The associations between the presentation modes and the test-takers' proficiency levels have been found in several previous studies (Buck, 1991; Chang, 2005; Chang & Read, 2013; Underwood, 1989; Wu, 1998; Yanagawa & Green, 2008). Although higher level L2 listeners' performance did not show a significant difference between the aural and written modes (Yanagawa

& Green, 2008) or was even better in the aural mode (Chang & Read, 2013), the majority of test-takers' perceived that items in written form were easier than those in the aural form (Chang & Read, 2006). Chang and Read (2013) reported that 53% of their higher proficiency level students responded that they did not have difficulties in reading and listening at the same time. They also discovered that in the written mode, as soon as the test-takers heard the answer that they thought correct, they would stop listening, move on to the next item, and read it beforehand. In the aural mode, however, it was impossible for them to move on to the next item, so they sometimes lost attention.

It is notable that the effect of presentation modes on the performance of lower level test-takers has been controversial in previous studies (Chang, 2005; Chang & Read, 2013; Underwood, 1989; Wu, 1998; Yanagawa & Green, 2008). Compared to higher level listeners, it was clear that they were not good at forming anticipations of the input and performed more uninformed guessing, which means that they did not benefit as much as higher level test-takers from the written mode (Chang, 2005; Wu, 1998). Still, Chang and Read (2013) reported that the lower level test-takers performed significantly better with the written mode. Yanagawa and Green (2008) also stated that less proficient L2 learners were more disadvantaged when they

could not access the questions in advance. They were less able to build a meaningful representation of a situation from the input without the support of the question. These findings were in line with that of Underwood (1989), who suggested that the written questions was helpful for lower level students.

The relatively limited short-term memory capacity of lower level test-takers might also account for their higher scores in the written mode than in the aural mode. Chang and Read (2013) explained that students from the lower level group were less able to hold the question and options in short-term memory when they were presented in the aural mode. Some of the students reported in their post-test discussion that they sometimes just guessed randomly, because they forgot the answer options which were given aurally.

Buck (1991) and Chang and Read (2006), on the other hand, revealed that lower level test-takers did not benefit by the written mode. Unlike higher level test-takers who successfully used written questions to get the idea of what to listen for so that they could focus on the key words, lower level test-takers often failed to recognize the topic and key terms in the item stem and the answer options that were provided in written form. Lower level students in Chang and Read (2013) also reported in their interview that long options

in written form caused them to give up reading, or if they could not finish reading, they would just guess. They also tended to depend much on word recognition – being readily diverted by distracters with words that match the recording, or confused by the use of negatives. They often considered the written questions and options as a source of background knowledge about the recording and made wrong assumptions. Regardless of their poorer performance in the written mode, however, they preferred the written form to the aural one.

In summary, previous studies had mixed results in the direction of the effects of the question/option presentation mode, particularly regarding the written answer options. Although the written mode provided test-takers with contextual information about the listening stimuli before listening, it allowed them to just match some words from the listening to choose the answer and the process did not seem to reflect the real-life listening very well. The effect of the presentation mode was also affected by the proficiency level of the test-takers. While higher level test-takers were less influenced by mode difference, its influence on the lower level test-takers' test-taking process and performance was more critical.

Most aforementioned studies compared the aural and the written mode

mainly focusing on the written mode's previewing effect, because in their studies, the test-takers could not access to the question until they listen to the passage in the aural mode items. However, previewing questions is not a distinct characteristic of the written mode and can also be applied to the aural mode by inserting the question before the listening stimuli.

Therefore, in the current study, questions are presented before the listening stimuli in the aural mode to balance the two modes regarding the effect of previewing questions. The present study aims to compare and contrast two modes concentrating on the fundamental difference between the two modes, reading vs. listening, and discuss its implication on developing a valid listening tests. In addition, the results will be discussed separately according to test-takers' proficiency levels in order to add more explanation to the previous studies' relatively contradictory results on lower level test-takers' performance.

Another factor investigated in this study is the effect of item types, and previous studies related to this topic is discuss in the following section.

2.2. Item Types

Item type, as a characteristic of the test method, is also considered to have influence on the performance of test-takers (Berne, 1992; Wolf, 1993). Previous studies on listening comprehension test item types mostly focused on comparing different response formats (Berne, 1992; Cheng, 2004; Hansen & Jensen, 1994). The item types that were frequently examined for the studies were multiple-choice, cloze or open-ended questions. For example, as a part of her research on the role of different factors in L2 listening comprehension assessment, Berne (1992) compared multiple-choice, open-ended, and cloze test scores of university students who are native speakers of English studying Spanish as a foreign language. She found that the test item type significantly affected test-takers' L2 listening comprehension performance. Participants who received the multiple-choice items scored higher than those who received either the open-ended or cloze items. She attributed this result to the different skills that each item type is requiring, which was also suggested by Shohamy (1984) and Wolf (1993). For the multiple-choice items, test-takers only had to recognize the correct response, whereas the open-ended and cloze items required them to retrieve and produce the correct response.

Cheng (2004) also compared EFL college students' performances in

traditional multiple-choice (MC), multiple-choice cloze (MCC), and open-ended (OE) questions and revealed that students performed significantly better in the MC and MCC questions than in the OE questions. Cheng attributed this result to the possibility of guessing and the availability of clues for the topic prediction in the MC and MCC questions and memory constraints in the OE questions. The students' perception on each item type also varied, indicating that most students preferred the selected-response types (MC and MCC) to the constructed-response type (OE) because they were less anxious when provided with the selected response questions.

In short, research on the effect of item types has been mainly on the comparison between different response formats, such as multiple-choice and open-ended. The discrepancy in the test results among the formats were reported to be largely due to the different skills that they are requiring. Although all the formats intended to measure the same construct, the listening ability, different formats could cause other factors to affect the performance. This suggests the need for further research on different item types within one response format, because even the same response format has several different item types that measure the same construct but require different skills. The current study specifically focuses on two different listening comprehension multiple-choice item types, dialogue-completion and Q&A, which are the two

most frequently used item types for the multiple-choice listening comprehension questions. These two item types are developed to measure the same construct, the listening ability, but require test-takers of different skills. The dialogue-completion type asks the test-takers to complete a short conversation by choosing the most appropriate response to the last turn of the provided listening stimuli, whereas the Q&A type requires them to get the main idea or make inferences based on the conversation. Therefore, the two item types of MCQ items might differently affect the test-takers' process and performance on the listening test.

Also, the earlier studies mentioned above had limitations in that they only employed listening comprehension items in written mode, which means that all the questions and options of the items used for their study demanded test-takers' reading ability, because they were written on the test paper. For example, Cheng (2004) clearly stated as one of the limitations of her study that all of her questions, regardless of item types, required students' reading skills as well as their listening skills. This gave rise to necessity of future studies on the varying effects of aural and written modes of question presentation when comparing different item types. The potential interaction between the presentation mode and the item type is discussed in the following section.

2.3. Potential Interaction between the Presentation Mode and the Item Type

Among the different studies on the effect of presentation mode discussed in Section 2.1, no studies on the effect of aural and written mode of question/option presentation made any distinction between different item types (Chang, 2005; Chang & Read, 2013; Wu, 1998; Yanagawa & Green, 2008). Chang and Read (2013), for example, did include different item types for their listening test, but did not mention which item types they used and did not report their effect on the results.

Considering item types when examining the effect of presentation mode in multiple-choice items is worth investigating, because combined characteristics of test method affect test-takers' cognitive processing and test scores in a dissimilar way (Bachman & Palmer, 1996). In other words, since different characteristics of test methods could interact with other characteristics, the impact of any single item characteristic on test-takers' performance needs a detailed analysis in terms of its interaction with other factors (Brindley & Slatyer, 2002).

In fact, Cheng (2004) specifically called for further study that

considers both mode and item type, since most research on item types, including hers, used items presented only in the written mode. Regarding the effect of presentation mode that was discussed in the previous section, therefore, examining how the mode difference works in relation to different item types is needed to acquire a more comprehensive picture of these two factors' effect. The present study investigated the effect of modes of question/option presentation together with that of different item types of multiple-choice questions on test-takers' listening comprehension test performance and their perception.

The potential interaction between the two test method characteristics, the presentation mode and the item type, was investigated for the current study in addition to the effects of each characteristic. For a more detailed analysis, the proficiency levels of the participants were taken into consideration. The following chapter describes the methodology employed in the study.

Chapter 3.

METHODOLOGY

This chapter describes the methodology employed in the present study. Section 3.1 discusses the participants. Section 3.2 provides details on the instruments in terms of the listening comprehension tests and the post-test survey. The procedure is described in Section 3.3 and the data analysis is explained in Section 3.4.

3.1. Participants

One hundred and fifteen Korean college students who have been learning English as a foreign language for 10 years on average participated in the study. Participants were recruited through online bulletin boards of two colleges, one of which is located in Seoul and the other in Cheongju, Korea. They came from a variety of majors/departments, including nursing, public administration, English literature, chemistry, computer information, medicine, and architecture. The participants consisted of 64 freshmen, 17 sophomores, 11 juniors, and 23 seniors.

The 115 students were divided into three groups of 38, 43, and 34 students according to their proficiency levels: the advanced level group with 700 points or above on TEPS[1] (or 850 or above on TOEIC[2])[3], the mid to high intermediate level group with 500 to 700 points on TEPS (or 650 to 850 on TOEIC), and the low intermediate level group with 500 points or lower on TEPS (or 650 or lower on TOIEC), respectively. If someone gets 700 or above on TEPS, one is considered to have advanced level of communicative competence, receiving a score of 2+ (The TEPS Council, 2009). According to TEPS Council (2009), a test taker with 2+ proficiency "will be able to do general tasks in English with a short, intensive training period" (p. 9). Test-takers who obtain 500 to 700 points, or a 2 or 3+, on TEPS are predicted to have a mid to high intermediate level of communicative competence. The TEPS Council explains that this level of test takers "will be able to do general tasks in English with a medium-length to long, intensive training period" (p.9). Finally, test-takers with 500 points or below on TEPS (3 or below) are

[1] Test of English Proficiency developed by Seoul National University

[2] Test of English for International Communication

[3] The conversion between TEPS and TOEIC scores is done following the conversion table provided by the TEPS Council (http://www.teps.or.kr/).

described to have low-intermediate level of listening competence. The TEPS Council depicts a test-taker with a score of 3 as “minimally able to do limited tasks in English with a medium-length to long, intensive training period” (p. 9).

3.2. Instruments

In this section, the instruments used for the present study are described including the listening comprehension test and a post-test survey.

3.2.1. The Listening Comprehension Test

The two item types selected for the present study are the dialogue-completion type and the question-and-answer (Q&A) type shown in Table 3.1. For the dialogue-completion task, test-takers were asked to complete a dialogue between two people. They listened to a short conversational exchange between a man and a woman, and were asked to choose the most appropriate response for the last turn. For the Q&A segment, test-takers were told to listen to a dialogue and answer comprehension questions asking main idea, specific detail, or inference. These two item types require different

listening skills in that the test-takers have to complete a short conversation by choosing the most appropriate and spontaneous response for the dialogue-completion type items, while they have to get the main idea or make inferences based on the conversation for the Q&A type items. Test items of the dialogue-completion type are from Part 2 of the TEPS listening section, and Q&A type items are from Part 3 of the TEPS listening section. Since the stimulus material for the dialogue-completion tasks is short conversations between two people, only dialogues, not monologues, were used for the Q&A items in this study to keep the stimulus of the two item types the same. All test items used for this study were sample questions from the previous TEPS tests, taken from an official TEPS practice book, *1200 Official TEPS Items* published by the TEPS Council in 2015. Samples of each item type are shown in Table 3.2 (Note that this table is the same as Table 2.1; it is presented again for readers' convenience).

Table 3.1

Listening Comprehension Test Used for the Study

Item types	Description
Dialogue-completion	Conversational exchanges (A-B-A-**?**)
Q&A	Dialogues – comprehension questions

Table 3.2

Sample Items Used for Each Format

Item Type	Listening Stimuli (Presented aurally)	Question/option Presentation Mode (How it is presented on test paper)	
		Aural Mode	Written Mode
Dialogue-	1. M: How do I get to the library? W: Follow this road to the big brick building. M: And that's the library? W: _____	1. (a) (b) (c) (d)	1. Choose the most appropriate response to complete the conversation. (a) It's my book. (b)No, but I'll try. (c) Yes, that's it. (e) Study them first.
Q&A	2. M: Are you transferring to a different university? W: Yeah, to one that's closer to home. M: Oh? Which one? W: Scarborough University. It's located near my family and has my major. M: That sounds good. W: Yes, I'm happy about it.	2. (a) (b) (c) (d)	2. What is the main topic of the conversation? (a) A plan to change schools (b) The high cost of education. (c) A comparison of two universities (d) Programs at Scarborough University

All dialogue-completion items have the same questions ("Choose the most appropriate response to complete the conversation.") and the objective of the items is quite clear for test-takers. Since the test-takers know the question before listening to the conversation, it might have a similar effect of previewing the item stem before listening to the conversation. On the other hand, Q&A items have various questions, so when answering the Q&A items delivered in the aural mode, test-takers would have to listen to the conversation, not knowing what to listen for. This imbalance between the two types makes it impossible to say whether the difference between the scores from the two formats is due to the item type or the preview of the item stem.

To make the two item types comparable regarding the previewing effect, the item stems were inserted additionally before the listening stimuli in the aural mode of the Q&A items. In this way, the test-takers could first listen to the question and then the conversation, followed by the question and the answer options. This measure reflects the cognitive process of real life listening tasks, since we usually have goals or purposes when we listen to someone. It is also in line with previous studies which showed a positive effect of the item stem preview on students' listening test performance (Buck, 1991; Yanagawa & Green, 2008). The overview of the listening comprehension test used for the study is summarized in Table 3.3 below and

the full test is provided in Appendix A.

Table 3.3

Overview of the Listening Comprehension Test Used for the Study

Section	Number of Questions	Item Type	Presentation Mode
1	4	Dialogue-completion	Aural
2	4	Dialogue-completion	Written
3	4	Q & A	Aural
4	4	Q & A	Written

3.2.2. Post-test Survey

A survey was developed to explore the participants' perceptions toward each section of the listening test, looking into their perceived difficulty, face validity, preference, and some possible factors that might have affected them. See Appendix B for the complete form of the survey. Question 1 examined the perceived difficulty of each section of the listening comprehension test and Question 2 examined their face validity. The face validity indicates the "appearance or perception of the test and how this may affect test performance and test use" (Bachman, 1990, p. 301). Question 3 and Question 4 were about preference questions, asking participants which mode

they prefer for each item type and why. Question 5-1 to Question 5-7 were developed based on Chang and Read (2013) to investigate factors that might have affected the item difficulty and validity. A five-point Likert response scale was used for each question except for Questions 3 and 4. For these two questions, only two options were provided since they asked participants' preference between the two presentation modes. Two open-ended questions were also employed as a part of Questions 3 and 4 to get participants' reasons for their choices.

3.3. Procedure

For the listening comprehension test papers, two forms, Form A and Form B, were developed to counterbalance the order of the sections. For dialogue-completion items, for instance, around one half of the test-takers from each proficiency group took Form A, receiving question 1-4 in the aural mode and 5-8 in the written mode. The other half of the test-takers who took Form B, on the other hand, received question 1-4 in the written mode and 5-8 in the aural mode. The two forms of the test and the form assignment for the three proficiency groups are shown in Table 3.4 and Table 3.5, respectively.

Table 3.4

Two Forms of the Test

	Form A	Form B
Dialogue-completion	Aural (1-4) – Written (5-8)	Written (1-4) – Aural (5-8)
Q&A	Written (9-12) – Aural (13-16)	Aural (9-12) – Written (13-16)
Total Items	16	16

Table 3.5

Form Assignment

	Low Intermediate (Group L)	Mid/high Intermediate (Group M)	Advanced (Group H)
Form A	21	15	18
Form B	22	19	20
Total	43	34	38

After taking their assigned listening comprehension test, participants completed the survey. Twelve participants were selected for the stimulated recall interviews, and they consisted of 4 participants from each proficiency group (approximately 10% of the number of participants in each group), 2 of whom took Form A and the other two who took Form B. The 12 participants

were asked to engage in stimulated recall immediately following their listening tests. Listening to the recording again, they reported how they chose the answer, what decisions they made in each step, and the reasons for their decisions. Also, some follow-up interview questions were given to ask participants to elaborate on their survey responses and to clarify their reasons. The verbal reports were conducted to qualitatively investigate their cognitive processes and factors that affected their judgements in each step of test-taking processes. All the processes of stimulated recall interviews were recorded and transcribed by the researcher.

3.4. Data Analysis

SPSS 22.0 for Windows was employed for the statistical analysis. For each proficiency group, a repeated measure two-way ANOVA was used for analysis to examine the effect of presentation mode and item type on test-takers' L2 listening performance and perception. Dependent variables were test-takers' test scores, perceived difficulty, or face validity, and independent variables were presentation mode and item type. A one-way ANOVA was also used to compare means of the three proficiency groups' responses for some survey questions.

Qualitative analyses for some post-test survey questions and the stimulated recall interviews were also conducted for an in-depth investigation of the meaning of the results drawn from quantitative analyses.

The data collection and the analysis were conducted based on the methodology described in this chapter, and the following chapter presents the results of the study.

Chapter 4.

RESULTS

This chapter presents the quantitative and qualitative results of the present study. Test-takers' performance and perceptions are reported in Section 4.1 and 4.2 respectively, and Section 4.3 provides an in-depth analysis of stimulated recall interviews.

4.1. Test-takers' Performance

For the first research question ("To what extent do mode of question/option presentation and item type affect three proficiency groups' L2 listening performance?"), a repeated measure two-way analysis of variance (ANOVA) was employed to examine the effect of mode and item type on the participants' scores on the listening comprehension test. The two independent variables were the item type (dialogue-completion and Q&A) and the presentation mode (aural and written). The dependent variable was the test-takers' performance (scores) on the test.

Table 4.1

Descriptive Statistics for Listening Comprehension Test Performance in Two Different Question Types and Two Different Modes

	DC-Aural		DC-Written		Q&A-Aural		Q&A-Written	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Group L	1.65	1.193	2.09	1.065	1.51	.935	1.70	1.103
Group M	2.97	.870	3.06	.814	2.53	.825	2.71	1.142
Group H	3.84	.370	3.82	.393	3.47	.687	3.58	.500

DC: Dialogue-completion type, Q&A: Question-and-answer type

The three proficiency groups' mean scores and standard deviations for four different sets of listening comprehension tests are presented in Table 4.1. One point was given to each item and the four formats, DC-Aural, DC-Written, Q&A-Aural, and Q&A-Written, had 4 items each, so the highest score and the lowest score one could get was 4 and 0, respectively. All groups obtained higher scores for the dialogue-completion items than for the Q&A items. For example, the mean score of Group L for the aural dialogue-completion items was 1.65, and it was higher than that for the aural Q&A items, 1.51. Group L's mean score for the written dialogue-completion items was 2.09, and it was also higher than its counterpart, written Q&A items, 1.70.

The similar tendencies appeared in the mean scores of the other two groups as well.

In terms of the presentation mode, Group L and Group M gained higher scores for the written mode than for the aural mode (aural = 1.65 and written = 2.09 on dialogue-completion, aural = 1.51 and written = 1.70 on Q&A for Group L; aural = 2.97 and written = 3.06 on dialogue-completion, aural = 2.53 and written = 2.71 on Q&A for Group M). However, this was not the case for the Group H, who obtained a slightly higher score for the aural mode (3.84) than for the written mode (3.82) on the dialogue-completion items. For the Q&A items, on the other hand, they received a higher score on the written mode (3.58) than on the aural mode (3.47).

Table 4.2 provides a summary of the repeated measure two-way ANOVA. Since no significant interaction effect between Presentation Mode and Item Type was shown in all three proficiency groups, the main effects of Mode and Item Type were analyzed. Significant main effects were shown for Item Type for Group M [F (1, 33) = 8.468, p < .006] and Group H [F (1, 37) = 11.426, p < .002], and their effect sizes were relatively large (partial η^2 = .204 for Group M and partial η^2 = .236 for Group H). However, no main effect was found for Mode for the two groups, which means that Group M

and Group H performed similarly on items presented in aural and written mode. For Group L, on the other hand, a significant main effect were detected for Mode [F (1, 42) = 4.394, p < .042, partial η^2 =.095], but no significant effect for Item Type.

Table 4.2

Results of the ANOVA for the Effects of Item Type and Mode on Test-takers' Listening Comprehension

	Source	*SS*	*df*	*MS*	*F*	*p*	partial η^2
Group L	Item Type	3.076	1	3.076	2.891	.096	.064
	Mode	4.238	1	4.238	4.394*	.042	.095
	IT * Mode	.703	1	.703	.738	.395	.017
	Error	40.047	42	.953			
Group M	Item Type	5.360	1	5.360	8.468**	.006	.204
	Mode	.596	1	.596	.766	.388	.023
	IT * Mode	.066	1	.066	.070	.793	.002
	Error	31.184	33	.945			
Group H	Item Type	3.480	1	3.480	11.426**	.002	.236
	Mode	.059	1	.059	.285	.597	.008
	IT * Mode	.164	1	.164	.635	.431	.017
	Error	9.586	37	.259			

To sum up, for the Groups M and H, the presentation mode did not make a significant difference, but the item type did. In both groups test-takers performed better with the dialogue-completion items than with the Q&A items. For the Group L, the presentation mode, not the item type, had a significant effect; the low-level students performed better with the written mode.

4.2. Test-takers' Perception

To answer the second research question ("What is the three proficiency groups' perception of aural and written modes of presentation on two item types?"), perceived difficulty and face validity of items in each format and test-takers' preference for the two presentation modes (aural and written) on the two item types (dialogue-completion and Q&A) were analyzed.

4.2.1. Perceived Difficulty

A repeated measure two-way ANOVA was used to investigate the impact of Mode and Item Type on the perceived difficulty of the items. The

independent variables were Mode (aural and written) and Item Type (dialogue-completion and Q&A), and the dependent variable was the perceived difficulty of the items in each format. The perceived difficulty of the items in each format was asked using the question "How easy or difficult was items in each format?" and a five-point Likert response scale was used for each question (e.g., '1' represents 'very easy' and '5' 'very difficult'). Means and standard deviations for three proficiency groups are shown in Table 4.3 below. In all three groups, the perceived difficulty was higher for the aural mode and the Q&A items than for the written mode and the dialogue-completion items, respectively.

Table 4.3

Descriptive Statistics for Perceived Difficulty of the Items in Two Different Question Types and Two Different Modes

	DC-Aural		DC-Written		Q&A-Aural		Q&A-Written	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Group L	2.62	1.188	2.24	.850	3.31	1.047	2.86	.899
Group M	2.88	1.094	2.41	.988	3.59	.892	3.06	.919
Group H	2.47	1.059	2.24	1.149	3.32	.842	2.89	.981

The results of the ANOVA on the perceived difficulty are summarized in Table 4.4. No interaction effect was found between Item Type and Mode. There was a statistically significant main effect for Mode in all three groups' perceived difficulty, which was significantly higher for the aural mode than the written mode of question and option presentation [F (1, 41) = 11.978, p < .001 for Group L; F (1, 31) = 15.583, p < .000 for Group M; and F (1, 37) = 8.388, p < .006 for Group H]. The effect size for Group M (partial η^2 = .321) was larger than Group L (partial η^2 = .226) and Group H (partial η^2 = .185), meaning that Group M's perception on the difficulty of the listening test was more affected by the presentation mode than that of other groups was. A significant main effect for Item Type was also revealed in all three groups' perceived difficulty, suggesting that the Q&A item type was perceived to be much more difficult than the dialogue-completion type in all three groups [F (1, 41) = 15.626, p < .000, partial η^2 = .276 for Group L; F (1, 31) = 15.127, p < .000, partial η^2 = .314 for Group M; and F (1, 37) = 23.347, p < .000, partial η^2 = .387 for Group H].

Table 4.4

Results of the ANOVA for the Effects of Item Type and Mode on Perceived Difficulty of the Listening Comprehension Items

	Source	*SS*	*df*	*MS*	*F*	*p*	partial η^2
Group L	Item Type	18.006	1	18.006	15.626***	.000	.276
	Mode	7.292	1	7.292	11.978**	.001	.226
	IT * Mode	.054	1	.054	.166	.685	.004
	Error	13.196	41	.322			
Group M	Item Type	15.559	1	15.559	15.127***	.000	.314
	Mode	8.500	1	8.500	15.583***	.000	.321
	IT * Mode	.029	1	.029	.063	.804	.002
	Error	15.471	33	.469			
Group H	Item Type	21.375	1	21.375	23.347***	.000	.387
	Mode	4.112	1	4.112	8.388**	.006	.185
	IT * Mode	.322	1	.322	1.505	.228	.039
	Error	7.928	37	.214			

4.2.2. Face Validity

To analyze the effect of Mode and Item Type on the face validity of the items in each format, a repeated measure two-way ANOVA is used. The independent variables were Mode (aural and written) and Item Type (dialogue-completion and Q&A), and the dependent variable was the face validity of the items. The face validity of the items in each format was asked with the question "How well do you think does each item format assess your listening competence?" and a five-point Likert response scale was used for each question (e.g., '1' represents 'very poorly' and '5' 'very well'). Table 4.5 presents means and standard deviations for the three groups' responses on the face validity of the items in each format.

Table 4.5

Descriptive Statistics for Face Validity of the Items in Two Different Question Types and Two Different Modes

	DC-Aural		DC-Written		Q&A-Aural		Q&A-Written	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Group L	3.47	.882	3.33	.865	3.65	.870	3.86	.833
Group M	3.62	.739	3.32	.768	3.91	.621	3.79	.770
Group H	3.39	.974	3.26	.891	3.84	.973	3.76	.820

For the dialogue-completion items, the face validity was higher with the aural mode than with the written mode in all three groups (Group L: aural = 3.47 and written = 3.33; Group M: aural = 3.62 and written = 3.32; and Group H: aural = 3.39 and written = 3.26). On the other hand, a different trend appeared among the different proficiency groups for the Q&A items. The face validity of the Q&A items was higher with the aural mode in Group M (aural = 3.91 and written = 3.79) and Group H (aural = 3.84 and written = 3.76), whereas it was lower with the aural mode than with the written mode in Group L (aural = 3.65 and written = 3.86).

According to the result of ANOVA, as summarized in Table 4.6, a marginally significant interaction effect between Mode and Item Type was detected in Group L [F (1, 41) = 3.941, $p < .054$], indicating that the participants with lower proficiency felt that the aural mode was more appropriate for the dialogue-completion items, while the written mode was more valid for the Q&A items. For the two other more proficient groups, Group M and Group H, no significant interaction effect was detected, and the two groups thought that the Q&A items can assess their listening ability much better than the dialogue-completion items [F (1, 31) = 5.555, $p < .025$ for Group M; and F (1, 37) = 9.715, $p < .004$ for Group H]. In terms of the presentation mode, the means of face validity were higher for the aural mode

in both dialogue-completion and Q&A, but there was no significant main effect.

Table 4.6

Results of the ANOVA for the Effects of Item Type and Mode on Face Validity

	Source	*SS*	*df*	*MS*	*F*	*p*	partial η^2
Group L	Item Type	5.587	1	5.587	9.515**	.004	.185
	Mode	.052	1	.052	.109	.743	.003
	IT * Mode	1.308	1	1.308	3.941	.054	.086
	Error	13.942	42	.332			
Group M	Item Type	4.971	1	4.971	5.555*	.025	.144
	Mode	1.441	1	1.441	2.788	.104	.078
	IT * Mode	.265	1	.265	.946	.338	.028
	Error	9.235	33	.280			
Group H	Item Type	8.526	1	8.526	9.715**	.004	.208
	Mode	.421	1	.421	1.238	.273	.032
	IT * Mode	.026	1	.026	.054	.817	.001
	Error	17.974	37	.486			

4.2.3. Mode Preference for Each Item Type

Participants were asked to choose one presentation mode they prefer for each item type. The reasons for their preference were elicited by open-ended questions following each preference question, and by seven five-point Likert response scale questions. The number and the percentage of the three groups' preference are shown in the Table 4.7.

Table 4.7

Mode Preference for Dialogue-completion and Q&A Item Types

Item Type	Mode	Group L	Group M	Group H	Total
DC	Aural	12 (27.9%)	10 (29.4%)	6 (15.8%)	28 (24.3%)
	Written	31 (72.1%)	24 (70.6%)	32 (84.2%)	87 (75.7%)
	Total	43 (100.0%)	34 (100.0%)	38 (100.0%)	115 (100.0%)
Q&A	Aural	9 (22.0%)	6 (18.2%)	5 (13.2%)	20 (17.9%)
	Written	32 (78.0%)	27 (81.8%)	33 (86.8%)	92 (82.1%)
	Total	41 (100.0%)	33 (100.0%)	38 (100.0%)	112 (100.0%)

All three groups answered that they preferred the written mode to the aural mode for both of the question types, dialogue-completion (Group L: Aural 27.9%, Written 72.1%, Group M: Aural 29.4%, Written 70.6%, Group

H: Aural 15.8%, Written 84.2%) and Q&A (Group L: Aural 22.0%, Written 78.0%, Group M: Aural 18.2%, Written 81.8%, Group H: Aural 13.2%, Written 86.8%). Also, more participants preferred the written mode of question/option presentation for the Q&A type than for the dialogue-completion type.

Responses they gave for the open-ended questions each of which immediately followed the two preference questions were categorized according to the participants' different reasons for the preference. Table 4.8 presents four different reasons for preferring the aural mode and thirteen different reasons for preferring the written mode in the dialogue-completion item type and in the Q&A item type, respectively. The frequencies of each response are also provided in the table. Since the participants chose the mode they preferred for each item type and gave their reasons for their preferences, the sum of the test-takers for each proficiency level in each mode is the same as the number of the test-takers who reported that they prefer the mode in Table 4.7. The difficulties in reading quickly and switching mode were the two main reasons participants did not like the written mode. The three most frequent reasons for preferring the written mode were memory constraints, the availability of prediction in the written mode, and the transient nature of speech in the aural mode.

Table 4.8

Reasons for Preferring the Aural Mode or the Written Mode

Preferring Mode	Reason for the preference	Dialogue-completion			Q&A		
		Group L	Group M	Group H	Group L	Group M	Group H
Aural	Reading in time was difficult.	5	5	0	2	4	0
	It was difficult to switch between modes.	1	4	3	1	2	0
	Aural mode is more valid.	4	1	4	5	0	4
	Aural mode was easier to concentrate.	2	0	0	0	0	0
Written	I could not remember all the options in the aural mode. It was confusing.	8	3	5	9	7	9
	I could predict the content of the listening stimuli before listening by reading the options.	7	7	5	7	7	7
	It is easy to miss questions and options in the aural mode.	6	1	7	5	1	7
	Reading is easier for me than listening.	3	5	5	3	6	3
	I could guess the answer by reading the options even if I had not understood the listening stimuli.	3	2	3	6	2	2
	Not having written options makes me nervous.	2	3	1	1	2	1
	I could read the options again and compare them easily.	0	2	1	0	1	2
	It is easier to cross out the wrong options one by one.	0	0	2	2	0	0
	I could have some time to think about what I just listened before choosing the answer.	0	0	0	1	1	1
	I didn't have to listen to the end once I found the answer.	0	1	1	0	0	0
	I am more used to written options.	0	0	1	0	1	0
	Listening questions and options are not appropriate for the test.	1	0	0	0	0	0
	I liked the written mode because it was more challenging.	1	0	0	0	0	0

Test-takers' perceptions on why each mode was easy or difficult were measured with seven questions using a five-point Likert response scale (e.g., '1' represents 'strongly disagree with the statement' and '5' 'strongly agree with the statement'), following Chang and Read (2013). The seven questions were:

Q5-1 I felt aural MCQ was easy because I did not have to worry about not understanding the questions and answer options;

Q5-2 I felt written MCQ was easy because I did not have to worry that I might not aurally comprehend the aural questions and answer options;

Q5-3 I felt aural MCQ was difficult because I could not read the questions and options before hearing the input;

Q5-4 I felt written MCQ was difficult because I could not finish reading the questions and options;

Q5-5 I had no difficulty remembering the questions and options while doing the aural MCQ questions;

Q5-6 Doing written MCQ was difficult because I had to read and listen at the same time;

Q5-7 When doing aural MCQ items, I did not have to wait until the speaker finished all the options. I chose the right one once I heard it.

Table 4.9

Descriptive Statistics for Seven Questions on Reasons for Mode Difficulty

	Source	Q5-1	Q5-2	Q5-3	Q5-4	Q5-5	Q5-6	Q5-7
Group L	Mean	2.52	3.60	3.33	2.81	2.48	2.57	2.86
(N=42)	SD	1.042	.939	1.162	1.174	.994	1.272	1.299
Group M	Mean	2.58	4.12	3.06	2.36	2.73	2.42	2.85
(N=33)	SD	1.324	.696	1.144	1.084	1.098	1.200	1.176
Group H	Mean	2.79	3.97	3.17	2.05	2.79	2.16	2.89
(N=38)	SD	1.212	.822	1.152	1.089	1.069	1.079	1.226

Means and standard deviations for the seven questions are shown in Table 4.9. Compared to the other two groups, Group H felt the aural mode much easier, because they could understand the questions and options well which were given aurally (Q 5-1). Group L's response for the questions 5-2 and 5-4 showed that they were less confident about reading and understanding the written questions and options than the other two groups were. Group L reported that they experienced difficulty when the written questions and options were not provided before the listening the input (Q 5-3) and when they have to remember the aurally given questions and options (Q 5-5). The higher the proficiency level was, the more the test-takers felt they did not have

difficulty in reading and listening at the same time (Q 5-6). For Question 5-7, three groups' responses were very similar (Group L: 2.86; Group M: 2.85; Group H: 2.89).

Table 4.10

Results of One-way ANOVA for the Seven Questions

		SS	*df*	*MS*	*F*	*p*
Q5-1	Between Groups	1.537	2	.768	.546	.581
	Within Groups	154.853	110	1.408		
	Total	156.389	112			
Q5-2	Between Groups	5.658	2	2.829	4.062*	.020
	Within Groups	76.608	110	.696		
	Total	82.265	112			
Q5-3	Between Groups	1.420	2	.710	.534	.588
	Within Groups	146.350	110	1.330		
	Total	147.770	112			
Q5-4	Between Groups	11.603	2	5.802	4.624*	.012
	Within Groups	138.007	110	1.255		
	Total	149.611	112			

Q5-5	Between Groups	2.202	2	1.101	.998	.372
	Within Groups	121.337	110	1.103		
	Total	123.540	112			
Q5-6	Between Groups	3.468	2	1.734	1.228	.297
	Within Groups	155.399	110	1.413		
	Total	158.867	112			
Q5-7	Between Groups	.045	2	.022	.015	.986
	Within Groups	168.964	110	1.536		
	Total	169.009	112			

The results of the one-way ANOVA of the three groups for each question are summarized in Table 4.10. According to the table, the difference between the three groups was confirmed to be significant in Q5-2 [F (2, 110) = 4.062, $p < .020$] and Q5-4 [F (2, 110) = 4.624, $p < .012$]. The post hoc comparison (Tukey HSD) showed significant differences between Group L and Group M for Q5-2 ($p < .021$) and between Group L and Group H for Q5-4 ($p < .009$). In other words, Group L was much more anxious about not being able to comprehend questions and options that are delivered aurally than Group M. Also, Group L had much more difficulty in reading the options in

allocated time with the written mode of question/option presentation, compared to Group H.

4.3. Stimulated Recall Interviews

In the stimulated recall interview, 12 participants (4 participants from each group) were asked to explain their thought processes on answering each item. The analysis of the stimulated recall interview data suggested some possible factors related to the two presentation modes (the aural and the written) on each item type (the dialogue-completion type and the Q&A type) that were found to affect the process and the performance of the participants in the listening comprehension test. The factors were categorized into two groups: factors associated with the characteristics of the aural mode (section 4.3.1) and those of the written mode (section 4.3.2). How these factors work differently in the two item types is examined and explained in each section. The factors are categorized and presented by the presentation mode, not the item type, to effectively show how the test-takers performed differently in each mode and whether and how these presentation mode effects are affected by the item type. The interview processes were done in Korean, and all Korean utterances in Examples were translated into English.

4.3.1. Issues Found in the Aural Mode

Two participants from Group L and three participants from Group M said they missed options while answering some questions in the aural mode. This phenomenon appeared in both dialogue-completion and Q&A types.

> *Example 1: Aural, Dialogue-completion, Participant L1*
> I knew (a) wasn't the answer and I missed (b), so I thought (c) was the answer. (The correct answer was (b).)
>
> *Example 2: Aural, Q&A, Participant L4*
> I heard (a) but I didn't hear (b), (c), and (d) well. So I just chose (a) for the answer. (The correct answer was (d).)

Most of the test-takers found it hard to remember all the options presented aurally and then to choose the best answer among them. Some participants mentioned that this problem was more serious in Q&A items, as can be seen in the Example 4, because Q&A items usually required more logical thinking than dialogue-completion questions.

> *Example 3: Aural, Dialogue-completion, Participant L4*
> By the time (c) and (d) came out, I forgot what (a) and (b) were. I just had to choose the answer randomly.
>
> *Example 4: Aural, Q&A, Participant M3*
> I didn't have any memory problem in dialogue-completion questions, but for Q&A questions, I often missed (c) and (d). I like to read options before listening but for these questions I couldn't, and I couldn't remember all the options.

Answering questions presented aurally was also affected by the order of the answer options. If the correct answer came first, participants often got confused by listening to other remaining options that followed the correct one, as illustrated in Example 5.

> *Example 5: Aural, Q&A, Participant H4*
> When I listened to (a), I thought it was correct. However, I continued listening to other options to be more certain, and this made me more confused. After listening to all the options, I was like "what was (a)?" and the options were all blurred in my head.

Some participants from Group H managed to reduce the memory problem by taking notes of all the options, as shown in Example 6.

> *Example 6: Aural, Q&A, Participant H3*
> I didn't have any problem with remembering the options, because I took notes of all the options as much as possible. Then I could easily compare the options.

In the written mode, the participants rarely made complete random guesses. They at least provided some pieces of evidence for their choices. In comparison, several random guesses were noted in the aural mode. Two participants from Group L and three participants from Group M showed random guessing for the answer in the aural mode, without having any clue from the listening. Example 7 and Example 8 demonstrate this.

Example 7: Aural, Dialogue-completion, Participant M4
I understood the conversation, but I couldn't find the answer among (a), (b), (c), and (d). So I just randomly picked one.

Example 8: Aural, Q&A, Participant M3
I didn't know the answer, so I just chose (b) because there seemed less (b)s than other options on my answer sheet.

In short, participants reported that they often missed options and found it hard to remember them when the options were presented aurally. Radom guessing was more frequently reported for the items in the aural mode. The participants from Group L and Group M, who have lower proficiency levels, had more of these problems compared to those with higher proficiency level, Group H.

4.3.2. Issues Found in the Written Mode

Reading answer options was an additional burden to some participants, particularly for Group L and Group M, while Group H did not have any difficulty with it. Example 9 and Example 10 show difficulty arising from having to read the options in the allotted time in the written mode. Example 11 illustrates a different tendency showed in a case of a participant from Group H who had relatively high reading ability.

Example 9: Written, Dialogue-completion, Participant L2
I think the pauses between the questions were too short to read all the written answer options. I didn't have time to read the options before listening to the text, so I started reading the options after the listening. Well, I managed to read all the options but there was little time left to understand them. I think these questions (in the written mode) were difficult because I didn't have time to read before listening.

Example 10: Written, Dialogue-completion, Participant M4
I had to choose the answer without fully understanding them because there was not enough time.

Example 11: Written, Q&A, Participant H3
I could read all the answer options before listening to the passage. There was enough time. Actually the pauses between questions were too long.

Mode switch between reading and listening was another difficult or annoying factor for some test-takers, as can be seen in the accounts below.

Example 12: Written, Q&A, Participant L3
The conversation was long, and I also had to read the options at the same time. This made me more confused, and I got lost.

Example 13: Written, Q&A, Participant M3
The listening started while I was still reading the options, so I missed the first part of the listening.

Example 14: Written, Dialogue-completion, Participant H2
It was the last line of the conversation that I have to find for the

answer, but I suddenly had to read the lines instead of listening and it wasn't that comfortable.

On the other hand, test-takers could take advantage of written options in several ways. First of all, many of them reported that they could predicted the topic of listening stimuli by reading the options in advance. This was especially helpful for those with high proficiency, as shown in Example 15, because their reading speed was high enough for reading all the options before listening to the passage.

Example 15: Written, Q&A, Participant H4
I read through the options before the listening came out, and then I could know that the listening would be about "leather jacket" and particularly "fake one." And then I heard "moral" and "cruelty" in the listening, so I immediately knew that (d) was the answer.

Employing this strategy on Q&A type items was reported to benefit participants more than doing so on dialogue-completion type items. Participant M1 explained that since the options in Q&A items were statements about the listening stimuli, she could predict the topic before actually listening to it.

Example 16: Written, Q&A, Participant M1
I read all the answer options before listening to the passage. I

> didn't do this for the dialogue-completion items because it was not really helpful to predict the listening stimuli, but it helped me a lot for the Q&A items.

However, it did not help participants from Group L much, because it was not possible for them to read options before listening to the passage, as shown in Example 17.

> *Example 17: Written, Q&A, Participant L3*
> I wanted to read the options before the listening, but I couldn't. I didn't have time to do so.

In addition, participants compared the options more frequently in the written mode than in the aural mode. Written options allowed them to reread and compare the options, crossing out the incorrect ones, while recorded options did not. Participant H2 provided more details about this process in Example 18.

> *Example 18: Written, Q&A, Participants H2*
> I first read from (a) to (d), and then looked for the answer, reading them again. I crossed out (a) and (b) which were certainly incorrect and I read (c) and (d) again because they were a bit confusing, and then chose (c).

Lastly, participants sometimes used word-matching strategy in the written mode. Without understanding the listening stimuli, they guessed the

answer by matching some words they heard from the listening stimuli to those in the written options, similar to what Participant L4 explained in Example 19. By catching the phrase "trying out," he chose the right answer (d), without understanding the conversation.

Example 19: Written, Q&A, Participant L4
I didn't understand the conversation, but I heard the phrase "tried out" so I chose (d) for the answer. (The correct answer was (d).)

Script:
W: How were the football tryouts?
M: The pickings were slim. It ended a lot earlier than I expected.
W: Not much athletic talent among the students?
M: I wouldn't know. They're not trying out.
W: Maybe they're discouraged by last year's state championship.
M: That was a fluke! They should be begging to join a team with a record as good as ours.

Question and answer options:
Q: What can be inferred from the conversation?
(a) The man is trying to recruit coaches for the football team.
(b) The team made a good showing at last year's state championship.
(c) The woman assisted the man with the football tryouts.
(d) Fewer people tried out for the team than the man expected.

This trend occurred more frequently in the Q&A section. However, this strategy was not always successful as shown in Example 20.

Example 20: Written, Q&A, Participant H1
Option (a) had exactly the same word "polite" which had

appeared in the listening stimuli, so I chose (a), not thinking deeply. (The correct answer was (d).)

In sum, written questions and options interfered with test-takers choosing the right answer in some situations, while they helped them in other situations. Mode switching in the written mode was confusing to a few participants regardless of their proficiency levels, and reading answer options were not easy for participants with lower proficiency. On the other hand, test-takers reported having taken advantage of written options by predicting the topic and using word-matching strategy particularly for Q&A items. The predicting strategy was exclusively used by the participants with relatively higher proficiency, while the word-matching strategy was often used by those with lower proficiency. The participants also compared options with ease in the written mode compared to the aural mode.

The participants' responses in stimulated recall interviews revealed some factors that made each mode difficult or easy for dialogue-completion and Q&A items by examining their test-taking process. Some factors were reported more frequently in specific proficiency groups and were more related to particular item type, while others were found in all

groups and both item types. These individual points that were described above will be comprehensively discussed in the following chapter.

Chapter 5.

DISCUSSION

In this chapter, the main findings are discussed with regard to the three research questions. Section 5.1 and 5.2 discuss the effects of mode and item type on the test-takers' performance and perceptions, and Section 5.3 examines factors which might have contributed to the effects. Lastly, Section 5.4 summarizes the previous sections and relates the results from the statistical analyses to the stimulated recall interview data.

5.1. The Effects of Mode and Item Type on the Test-Takers' Performance

For the first research question ("To what extent do mode of question/option presentation and item type affect three proficiency groups' L2 listening performance?"), the results varied according to different proficiency groups. The low intermediate group, Group L, performed significantly better in the written mode, while they did not show any significant difference for item type. On the other hand, the mid/high

intermediate level group and the advanced level group, Group M and Group H, attained similar scores for both modes, while they significantly scored higher for the dialogue-completion items than the Q&A items.

This indicates that the higher proficiency groups performed equally well whether the questions and answers were recorded or written. This result is consistent with earlier studies which compared the scores of aural and written mode between participants with higher and lower proficiency levels (Chang & Read, 2013; Yanagawa & Green, 2008). The two higher proficiency groups were rather influenced by the item type, scoring higher on the dialogue-completion items. Relatively lower scores on the Q&A type items were somewhat expected due to its longer input and options that could have made it more difficult. While the input listening stimuli for the dialogue-completion items consisted of 3 turns, that for the Q&A items consisted of 6 turns. The answer options for the Q&A items were relatively longer than those for the dialogue-completion items. The average number of words in an option for a Q&A item was 8.1, whereas that for a dialogue-completion item was 6.2. The longer listening stimuli and options might have adversely affected the participants' performance.

The lowest proficiency group performed differently. Although they did feel that the Q&A items were much more difficult than the dialogue-completion items (see Section 4.2), the result shows that they obtained similar scores on both item types. However, they were significantly influenced by the presentation mode, scoring much lower when the questions and options were given aurally. This could mean that the difficulty they had in listening to all four options and processing them far outweighs that in reading the options and processing them in time. Therefore, choosing between the aural and written mode in the development process of a listening test may have much more critical influence on the test-takers with the low-intermediate proficiency level than on any other groups with higher proficiency, whether the items are dialogue-completion or Q&A.

5.2. Test-Takers' Perception of Aural and Written Modes of Presentation on Two Item Types

For the second research question about the perceived difficulty ("What is the three proficiency groups' perception of aural and written modes of presentation on two item types?"), relatively consistent results for the

perceived difficulty were obtained throughout the three different proficiency groups. All three groups felt that the aural mode was more difficult than the written mode. Chang and Read's (2006) study also had the same result, in which test-takers felt the written mode easier than the aural mode. In terms of the item type, all groups thought that the Q&A type items were more difficult than the dialogue-completion type items.

A different trend appeared for the face validity, which shows the test-takers' perception on how well the test measures their listening ability. Group M and Group H's responses revealed that the face validity was significantly higher for the Q&A items which they said were more difficult. However, although the face validity of the aural mode was higher than that of the written mode for Group M and Group H, the differences were not significant. Group L's response was different from those of the other two groups. There was a marginally significant interaction effect of Mode and Item Type for Group L, suggesting that the lowest proficiency group somewhat felt that the aural mode was more appropriate for the dialogue-completion items and the written mode was more suitable for the Q&A items.

The result that the Group L was the only group who felt that the written mode can assess their listening ability better than the aural mode in a listening

test for at least one item type might have a close relation to the Group L's performance discussed in the previous section. Group L was the only group whose listening comprehension scores were significantly lower in the aural mode than in the written mode. This trend existed in the earlier studies such as Chang and Read (2013) and Yanagawa and Green (2008), which revealed that the lower level test-takers were more disadvantaged without written items. A possible explanation for these results is that among the L2 listeners who were reported to have limited short-term memory when listening in the L2 (Cook, 2013; Ohata, 2006), participants with lower proficiency were even more affected by the memory problem associated with the aural mode. This possibility was also suggested in Chang and Read (2013). Also, without the written options, they could not use the word-matching strategy that they frequently resorted to when they could not fully understand the listening stimuli. A more detailed and comprehensive examination on the factors that affected the participants' performance and perception is discussed in Section 5.3.

When asked which mode they prefer, the majority of all three groups answered that they preferred the written mode to the aural one. The positive attitude toward the written mode was comparable to those reported by Buck (1991) and Chang and Read (2006). It was interesting to note that there was

a tendency for the higher proficiency group to prefer the written mode more strongly. Even though Group H did equally well on both modes of tests and was not affected much by the mode (see Section 4.1.), they strongly preferred the written mode for both dialogue-completion and Q&A items, and their preference was stronger than any other group.

The high proficient participants who preferred the written mode provided similar reasons for preferring the written mode to those provided by the other two groups' participants who preferred the written mode. The major three reasons were that there was no memory burden in the written mode, that they could predict the listening stimuli by reading the answer options, and that they did not have to worry about not hearing and missing any options because of the transient nature of the aurally given options. The difference between Group H and the other two groups lies in the reasons for choosing the aural mode. That reading the options in time was difficult was the major reason for the Group L and Group M's participants who preferred the aural mode. The participants from Group H who preferred the aural mode, however, did not have this reading problem. Not even a single person from Group H said that they preferred the aural mode because they could not read the questions and options fast enough. Their reasons for choosing the aural mode were that they thought the aural mode was more valid for assessing listening

and that it was somewhat uncomfortable and unnatural to switch modes from listening to reading.

This tendency was also supported by the results of some additional survey questions (Q5-1 to Q5-7) on why each mode was easy or difficult. It was noteworthy that Group H was more confident reading the questions and options than listening to them. When they were asked if they felt the written mode was difficult because they could not finish reading the questions and options, their response on the Likert scale was significantly lower than Group L. On the contrary, there was no significant difference between the two groups when they were asked if they felt the written mode was easy because they did not have to worry that they might not aurally comprehend the aural questions and options. To sum up, Group H's reasons for preferring the written mode indicated that not only did Group H not have strong need for recorded questions and options because they were perfectly comfortable with reading them on paper in time, but also they were more confident about reading than listening the questions and options.

5.3. Factors Found in the Different Test Formats and Their Effects on Test-takers' Performance and Perception

From the stimulated recall interviews on the test-takers' process of arriving at the answers in the listening comprehension test, some factors that were not directly relevant to the listening comprehension were found in both aural and written mode. For the aural mode, the need of a high level of concentration and good memory seemed to be the most problematic for test-takers. Once the recorded questions and options came out, they disappeared. Test-takers therefore needed to sustain a high level of concentration all the time not to miss any option just because they lost their concentration for a couple of seconds.

This transient nature of the aural mode also required good short-term memory, which was reported to be more limited for L2 learners (Cook, 2013; Ohata, 2006) and to be even more constrained for L2 learners with lower proficiency (Chang & Read, 2013). If participants wanted to compare options, they had to do it in their memory, recalling the options and the listening stimuli and comparing them at the same time. Some test-takers with high levels of proficiency managed to overcome this problem by taking notes for all of the questions and options, visualizing options like in the written mode.

However, the lower level participants did not employ the same strategy. Listening and writing at the same time might be another high level ability that the lower level participants did not possess.

This also led to another problem found in the aural mode, the random guessing of the answer. Wu (1998) noted uninformed guessing as one of the construct-irrelevant factors found in the multiple-choice (MC) questions in general, and MC questions presented aurally seemed to exacerbate this problem. For the aural mode items, participants from Group L and Group M often reported to have made random guesses about the answer without any clue. In comparison, the same participants did rely on some evidence when they had to make guess for the items in written mode. For the written mode, they could cross out the options that were most unlikely to be the answer and then had a guess at the answer among remaining options. For the aural mode, however, when they were not sure about the answer immediately, they often failed to recall all the options and ended up making random guesses.

For the written mode, the most salient one of the construct irrelevant factors was that it required certain level of reading ability, and the better a participant was in the reading comprehension, the more beneficial it was. This unfavorably affected the participants with lower proficiency, because their

reading skills and speed were often not as good as those that are required for understanding the written options in time. The participants from Group L frequently reported that they had difficulty reading the questions and options and comprehending them quickly. They felt that the time provided for reading was not enough, while the participants from Group H felt that the time provided was so long that they even thought that it was too boring.

Mode switching was another factor that was required in the written mode items. The test-takers had to listen to the passage and read the questions and options alternately or simultaneously. It was also more disadvantageous for the lower level participants. They were sometimes confused by reading and listening at the same time, or missed the first part of the listening stimuli, because they were concentrating on reading the options. The confusion that the test-takers experienced was also reported by some previous studies (Ur, 1984; Weir, 1993), which found that test-takers' attention to listening was distracted by the written options.

Some factors found in the stimulated recall interviews were helpful to the test-takers in terms of finding the answer, but not directly relevant to listening comprehension. The written mode allowed the test-takers to read the options before listening to the passage. In this way, they were able to predict

the topic or general ideas of the listening stimuli in advance, a behavior which was also discovered in earlier studies (Buck, 1991; Yanagawa & Green, 2008). This strategy was particularly used for Q&A items by Group H participants and some Group M participants, indicating that their reading speed was fast enough to save some time for reading the options for the next question. Participants from Group L, on the other hand, were not able to apply this strategy due to their low reading speed. Most of them could not read the options beforehand, even though they wanted, so were not able to predict the topic or ideas about the listening stimuli. This result was in line with Wu's (1998) finding in his study on 10 Chinese ESL students' listening comprehension test-taking processes using retrospective verbal report. He also concluded that viewing the questions and options helped advanced students by allowing them to predict the listening, but it was not beneficial to students with lower English proficiency. Also, even if they managed to read the options before the listening, they could not effectively predict the idea or topic of the listening stimuli. The unsuccessful use of the written questions and options was also discovered by the findings of Buck (1991) and Chang and Read (2006).

Instead, the participants with lower proficiency frequently used the word-matching strategy. With the written options presented before them, they

caught some words from the listening stimuli and chose the answer option that matched the words they heard. The strategy was not always successful, but was often used when they could not understand the listening. The participants from Group H did not resort to this strategy very often, because they could understand the listening. It was interesting to note, however, that they also tried using the word-matching strategy, when they missed or did not understand some of the information. Again, even for the participants with higher proficiency, it was not easy to always get a right answer by using this strategy. This lexical attractiveness has been reported by several previous studies (Freedle & Fellbaum, 1987; Freedle & Kostin, 1996, 1999). Freedle and Fellbaum (1987) found overlapping words between single-sentence listening comprehension passage and the answer options played a significant role in determining the item difficulty, and Freedle and Kostin (1999) yielded a similar result using longer listening stimuli, TOEFL mini-talks. Yanagawa and Green (2008) also noted that previewing answer options encouraged students to use lexical matching strategy.

The results of participants' preference showed that all three groups showed a stronger preference for the written mode in Q&A items than in the dialogue-completion items (see 4.2.3.). This can also be partly explained by the participants' use of the two strategies: prediction and word-matching. The

result that the prediction strategy was more frequently used for Q&A items than dialogue-completion items means that the strategy was more useful for Q&A items particularly for the high level test-takers. The word-matching strategy was also more suitable for Q&A items in the written mode, which made the written mode more attractive for Q&A items for the lower level test-takers as well.

5.4. Summary and Further Discussion

The test-takers with mid/high intermediate to advanced proficiency level (Group M and Group H) and those with low intermediate level (Group L) performed and felt differently in the listening test that was given in the two modes and the two item types. The discrepancies were found in their scores, perceptions, and preferences, and factors affected these three aspects. Group L was significantly influenced by the mode, receiving much lower scores in the items in aural mode. This result was closely related to their perceived difficulty and the face validity of each format and their preference as well. They felt the aural mode was much more difficult than the written one, and preferred to have items with written questions and options. Unlike Group M and Group H who perceived the aural mode as a more valid version of

listening tests for both the dialogue-completion and Q&A types, Group L thought the aural mode was better for the dialogue-completion items whereas the written mode was more appropriate for the Q&A items.

The reasons Group L found the aural mode much more difficult were that it required them with a high level of concentration and sufficient short-term memory capacity, in addition to their listening skills. Using the aural mode in this sense might undermine the validity of the listening test. To some extent, however, Group L's higher score in written mode was partly due to their tendency to resort to the word-matching strategy for the items in written mode. This can negatively affect the validity of the test, because the test-takers just matched some words from the listening to the written options, not understanding them. In addition, reading options in time was an obstacle for taking the listening test in the written mode for many Group L participants. In other words, they could not get the item right if they could not finish reading the options in time, even if they understood the listening stimuli.

Group M and Group H's scores, on the other hand, were significantly higher for the dialogue-completion items than for the Q&A items. Unlike Group L, Group M and Group H were not significantly affected by the presentation mode of the questions and options. Still, they felt the aural mode

was much more difficult than the written mode, and Group H even expressed a stronger preference for the written mode than the other two groups did. Their favorable attitude toward the written mode could be explained by the reasons found in the survey and the interview. Even though they could perform equally well on both modes, they were more confident about reading the questions and options than listening them. Additionally, they did not have any difficulty in reading in time, which was not an easy task for Group L. With their high reading proficiency, they could read the options in advance and predict the listening stimuli before actually listening to it. Group L did not get this chance to predict the topic because their low reading proficiency prevent them from reading the options in advance. Thus, since the written mode inevitably requires a certain level of reading ability, the test-takers' reading ability, in addition to the listening ability, can critically affect the test-takers' scores.

All things considered, whether to use the aural or the written mode is not a simple decision in the listening test development process, because diverse factors are associated with different proficiency groups. For the listening tests that intend to measure listening ability, the memory capacity and the reading ability were found to be the two major construct irrelevant factors that play an important role in the aural and written mode, respectively.

The test-takers with lower proficiency were more critically affected by these factors both for the dialogue-completion items and the Q&A items. Therefore, since we do not want the memory capacity or the reading ability to decide the listening test results in most cases, the effort to minimize their effects is necessary when developing the tests.

The aural mode is more recommendable in that it does not require reading ability which could hinder the test-takers with low reading proficiency from fully demonstrating their listening ability and which allows word-matching strategy. Its drawback, that it requires good memory, can be alleviated by making the options simple and clear and by reducing the number of options from four or five to three. The aural mode may be particularly appropriate for the dialogue-completion item type, because it is a part of the conversation that the test-takers are looking for. In fact, even some participants with high proficiency sometimes found mode switching in the dialogue-completion items uncomfortable (see section 4.3.2.). It would not be natural to switch from listening to reading when they have to complete the conversation by choosing the most appropriate response.

Employing the written mode for the Q&A item type could be justified if the options are seen to be too long or difficult for test-takers to process

within their short-term memory only by listening them. The written mode can also have a positive effect on the test-takers by reducing their test anxiety. Yet, if the questions and options are to be delivered in written form, they should be written in easy language not to require a high level of reading ability. The most important point, however, is to take the target test-test takers' proficiency levels into consideration. If the target test-takers of a test tend to have a lower English proficiency level, the above discussions should be thoroughly reviewed and considered since test-takers with lower proficiency are much more susceptible to the mode difference than those with higher proficiency.

Chapter 6.

CONCLUSION

This chapter is composed of three sections. Section 6.1 summarizes the major findings of the present study. In Section 6.2, the implications are presented on construct validity of the two modes of listening tests. Finally, Section 6.3 reports the limitations of the present study and makes suggestions for the further research.

6.1. Major Findings

This study investigated the effect of question/option presentation mode and item type on Korean EFL learners' listening comprehension performance and their perception toward the different MCQ test formats. Regarding the first research question, test-takers with a low intermediate level of English proficiency performed significantly better in the written mode, while those at mid to high intermediate and advanced proficiency levels showed little difference in scores between the two modes. The lower proficiency group participants were more negatively affected by not having

the written questions and options, because they found it difficult to remember the options in the aural mode. Their particularly low scores in the aural mode might have been partly caused by preventing them from using the word-matching strategy, which they frequently resorted to in the written mode.

In terms of the second research question, test-takers from Group M and Group H perceived that the Q&A items were much more difficult but seemed more valid as listening test items than the dialogue-completion items. On the other hand, they felt that the aural mode was significantly difficult than the written mode, but there was no marked difference in the face validity between the two modes. Only the lowest proficiency group showed a marginally significant interaction effect between item type and mode for the face validity, showing that they felt the aural mode was more appropriate for the dialogue-completion items, whereas the written mode was better for the Q&A items. In other words, the participants with the lowest proficiency were the only group who thought that the written mode was more valid, at least for the Q&A items. They felt that they could not fully demonstrate their listening ability through the aural mode. In fact, they were the only group who received significantly lower scores in the aural mode than in the written mode.

In addition, all three groups preferred the written mode to the aural mode. The advanced group (Group H) showed a particularly strong preference toward the written mode. This was because they did not have any difficulty in reading the questions and the options in time and even in advance and were able to use them as clues for predicting the listening stimuli and the answer. Having the written questions and options was more beneficial to the advanced group than to the other two groups, mid/high intermediate and low intermediate groups, who had limited reading ability compared to the advanced group.

The difficulties that the test-takers reported in the survey on their perception were concretized by the stimulated recall interviews. Major construct irrelevant factors for the listening comprehension test found in the aural and written mode of question/option presentation were the memory capacity and the reading ability, respectively. The aural mode prevented the test-takers from having written options before them and comparing them. Therefore, the lower proficiency group, who are believed to have a shorter working memory in L2, suffered from missing and forgetting the answer options. This also increased their tendency to make random guesses for the answer without any logical thinking.

There were some construct-irrelevant factors associated with the written mode, all of which were closely related to the participants' reading ability. Most importantly, the test-takers with higher proficiency were able to read the questions and options fast enough, and were less confused by mode switching. This provided them with more time left to think again and compare the options. For Q&A items, they frequently read the options before listening to the passage and predicted the main ideas and topics of the passage. This can negatively affect the validity of the listening test, since the test-takers could sometimes predict the answer before even listening to the stimuli. The lowest proficiency group, however, could not take advantage of the characteristics of the written mode, since they were not fluent readers. Sometimes, they could not even finish reading the options in time. Thus, not only did their limited listening ability lead to low scores in the listening test but their low reading ability also kept them from performing to the best of their ability. They did try to use the written options when they could not fully understand the listening stimuli by employing the word-matching strategy. Although it was not always as successful as they expected, this strategy could harm the validity of the test, because the test-takers just matched a couple of words from the listening to the options without understanding the listening stimuli.

6.2. Implications

Since construct validity is "central to the appropriate interpretation of test scores" (Bachman, 1990, p. 255), identifying constructs that a test is measuring is one of the key issues in the test development process. This study has drawn attention to several issues related to developing and choosing listening comprehension multiple-choice tests regarding different formats.

Both the aural and the written modes are found to have strengths and weaknesses with respect to construct validity. The influence of reading ability on test-takers' listening comprehension performance could be avoided with the aural question/option presentation mode by providing all questions and options aurally, but this imposed an additional memory burden on the test-takers. On the other hand, test-takers felt less anxious with written questions and options, because they did not have to remember them. However, the written mode required a certain level of reading proficiency and entailed other reading related construct-irrelevant factors, such as the ability to predict the main idea or even the answer by only reading the options and the use of word-matching strategies. The impact of each mode's characteristics was intensified in Q&A items, since they had longer options compared to the dialogue-completion ones. Therefore, test developers and teachers should be

aware of the factors that might influence the outcomes in each mode and decide which mode to use for each item type, considering the target population.

The proficiency level of the target population also needs to be taken into consideration when making decisions on choosing or developing a listening comprehension test. Participants with lower proficiency were more affected by mode difference. Thus, if the target population of a listening test includes participants with low intermediate proficiency level or even lower levels, the test developers and teachers should keep in mind that some irrelevant constructs that they might not intend to measure, the memory capacity or the reading ability, could significantly affect the test-takers' listening performance.

Based on participants' performance and perception on the different formats of listening tests and reports in the recall interview, the two suggestions can be made for test developers and teachers who develop or choose listening comprehension multiple choice tests for EFL learners. First of all, it is most recommendable to give the questions and options aurally, because the written mode requires a certain level of reading ability and invites word-matching strategy. The memory burden on the test-takers in the aural

mode could be relieved by reducing the number of options and by making them short and clear. This is particularly advisable for the dialogue-completion items, since the test-takers have to choose a response that is a part of the whole aurally-given conversation.

However, the written mode can be more appropriate in some cases, such as when the options are too long for the aural mode and too difficult for the target test-takers to process with their memory capacity only by listening. If a listening test was to employ the written mode, the questions and options should be as simple and clear as possible. It should not include any difficult vocabulary or sentence structures that test-takers with lower reading ability cannot understand.

Again, it is the characteristics of the target test-takers and the item types that are the first thing to consider when deciding the presentation mode of questions and options in a listening test. Only by considering the target test-takers' short-term memory capacity and reading ability, and by scrutinizing the property of item types that the test will use, the test developers can figure out the best way to present the questions and options in their listening test.

6.3. Limitation and Suggestions

The results of this study have examined the effects of presentation mode in the two different item types on the test-takers' performance and perception, and investigated some possible reasons for the differences. However, there are some limitations in this research that could be improved and developed in the future study.

The most obvious limitation in this study was that of a small sample size for the stimulated recall interviews. The number of participants who took part in the interview might not have been representative enough to generalize the findings. Still, the interviews did reveal some important issues regarding the presentation mode and the item type for the listening multiple-choice questions.

Another limitation of the study is that participants' proficiency levels varied within each group. Participants from Group M were particularly diverse in their proficiency levels, therefore the performance of the group's higher end was more like the advanced level while that of the lower end was similar to the low intermediate level. This could be improved in the future research by making the distinctions between the groups more clear when forming them.

The findings and suggestions of the current study can be developed into future research regarding the following two research topics. Firstly, to relive the memory burden of the aural mode not by giving the written options but by reducing their number, the interaction effect between the number of options and the presentation mode needs to be investigated. Secondly, the appropriate length or complexity of the language for the options in the written mode can be examined in relation to the test-takers' proficiency levels. These further research would help provide a clearer picture about how to employ the aural and the written mode for the listening tests with multiple-choice questions.

Reference

Bachman, L. F. (1990). *Fundamental considerations in language testing*: Oxford, UK: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford, UK: Oxford University Press.

Berne, J. E. (1992, August). *The role of text type, assessment task, and target language experience in L2 listening comprehension assessment*. Paper presented at the the Annual Meetings of the American Association for Applied Linguistics and the American Association of Teachers of Spanish and Portuguese, Cancun, Mexico.

Berne, J. E. (1995). How does varying pre-listening activities affect second language listening comprehension? *Hispania, 78*(2), 316-329.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing, 19*(4), 369-394.

Buck, G. (1990). *The testing of second language listening comprehension.* Lancaster, UK: University of Lancaster.

Buck, G. (1991). The testing of listening comprehension: An introspective study. *Language Testing, 8*(1), 67-91.

Buck, G. (2001). *Assessing listening*. Cambridge, UK: Cambridge University Press.

Chang, C.-S. (2005). The perceived effectiveness of question preview in EFL listening comprehension tests. *New Zealand Studies in Applied Linguistics, 11*(2), 75-96.

Chang, C.-S. (2008). Listening strategies of L2 learners with varied test tasks. *TESL Canada Journal, 26*(1), 1-26.

Chang, C.-S., & Read, J. (2006). The effects of listening support on the listening performance of EFL learners. *TESOL Quarterly, 40*(2), 375-397.

Chang, C.-S., & Read, J. (2013). Investigating the effects of multiple-choice listening test items in the oral versus written mode on L2 listeners' performance and perceptions. *System, 41*(3), 575-586.

Cheng, H. F. (2004). A comparison of multiple-choice and open-ended response formats for the assessment of listening proficiency in English. *Foreign Language Annals, 37*(4), 544-553.

Cook, V. (2013). *Second language learning and language teaching*. London, UK: Routledge.

Freedle, R., & Fellbaum, C. (1987). An exploratory study of the relative difficulty of TOEFL's listening comprehension items. In F. Feedle &

R. Duran (Eds.), *Cognitive and linguistic analyses of test performance* (pp.162-192). Norwood, NJ: Albex.

Freedle, R., & Kostin, I. (1996). *The prediction of TOEFL listening comprehension item difficulty for minitalk passages: Implications for construct validity*. ETS Research Report Series No. RR-96-29. Princeton, NJ: Educational Testing Service.

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing, 16*(1), 2-32.

Hansen, C., & Jensen, C. (1994). Evaluating lecture comprehension. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 241-268). Cambridge, UK: Cambridge University Press.

Hemmati, F., & Ghaderi, E. (2014). The effect of four formats of multiple-choice questions on the listening comprehension of EFL learners. *Procedia-Social and Behavioral Sciences, 98*, 637-644.

Iimura, H. (2010). Factors affecting listening performance on multiple-choice tests: The effects of stem/option preview and text characteristics. *Language Education & Technology, 47,* 17-36.

Messick, S. (1996). *Validity and washback in language testing.* ETS Research Report Series No. RR-96-17. Princeton, NJ: Educational Testing

Service.

Nissan, S., DeVincenzi, F., & Tang, K. L. (1995). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension*. ETS Research Report Series No. RR-95-37. Princeton, NJ: Educational Testing Service.

Ohata, K. (2006). Auditory short-term memory in L2 listening comprehension processes. *Journal of Language and Learning, 5*(1), 21-28.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing, 1*(2), 147-170.

The TEPS Council. (2009). *TEPS*. Retrieved from http://www.teps.or.kr.

The TEPS Council. (2015). *TEPS choisin gichul 1200-jae [1200 official TEPS items]*. Seoul, Korea: Nexus.

Underwood, M. (1989). *Teaching listening.* Boston, MA: Addison-Wesley Longman Ltd.

Ur, P. (1984). *Teaching listening comprehension.* Cambridge, UK: Cambridge University Press.

Weir, C. J. (1993). *Understanding and developing language tests.* Upper Saddle River, NJ: Prentice Hall.

Weir, C. J. (2005). *Language testing and validation*. London, UK: Macmillan.

Wolf, D. F. (1993). A comparison of assessment tasks used to measure FL reading comprehension. *The Modern Language Journal, 77*(4), 473-489.

Wu, Y. A. (1998). What do tests of listening comprehension test?-A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing, 15*(1), 21-44.

Yanagawa, K., & Green, A. (2008). To show or not to show: The effects of item stems and answer options on performance on a multiple-choice listening comprehension test. *System, 36*(1), 107-122.

Appendix A. Listening Comprehension Test

Form A

<Dialogue-completion>

1-1 Aural Mode

1. (a) (b) (c) (d)

M: Are you ordering the hamburger or burrito?
W: I don't know; they both look good.
M: How about we share?
W: _____
(a) Good thinking. We can split one of each.
(b) I don't know what they serve.
(c) No, the meal is on me.
(d) Sure, we can get something else.

2. (a) (b) (c) (d)

M: Did you get everything on the grocery list?
W: Almost – they were out of cherries.
M: But I need those for the muffins.
W: _____
(a) OK, I'll take them out.
(b) Just don't include them.
(c) They didn't have those, either.
(d) No, that recipe needs cherries.

3. (a) (b) (c) (d)

M: I’m considering hiring a cleaning service.

W: Aren’t they expensive?

M: Yeah, but I’m too busy to keep my house tidy.

W: _____

(a) That’s much longer than I spend.

(b) Then maybe it’s worth the price.

(c) I usually keep mine at home.

(d) Ask for your money back.

4. (a) (b) (c) (d)

M: Professor, can I turn in my essay tomorrow?

W: You know it’s due today, right?

M: Sorry, I’ve been ill all week.

W: _____

(a) I’ll give it to you later.

(b) Sure, today works for me.

(c) Then I’ll make an exception just this once.

(d) I can’t return it today, anyway.

1-2 Written Mode

5.

W: How's your job searching going?
M: I've submitted several applications.
W: But no offers yet?
M: _____

(a) No, but I'm staying hopeful.

(b) I'll decide soon enough.

(c) Right. I don't know which to choose.

(d) I haven't applied there yet.

6.

W: Can you come to Friday's basketball game?
M: Hmm… Friday's not great for me.
W: Don't tell me you've already got plans!
M: _____

(a) I do, but I'll try to change them.

(b) Even so, I hope you'll come.

(c) I can make it if you move it to Friday.

(d) Honestly, I won't forget.

7.

W: Hi, I placed an online order but accidentally gave the wrong address.
M: OK, do you know if it's been shipped yet?
W: I doubt it. I placed it a minute ago.
M: _____

(a) You can just order it online.

(b) We apologize for the mistake.

(c) Then I can help you change it online.

(d) Sorry, the order hasn't been shipped.

8.

M: Can you give this file to Aaron when he gets back on Monday?
W: Sure. He's not here today?
M: He's on vacation. You didn't know he was gone?
W: _____

(a) No, he already asked me about it.

(b) Still, I'll be glad to have him back.

(c) It has been a while since I told you.

(d) I've been so busy that I didn't notice.

<Q&A>

2-1 Aural Mode

9. (a) (b) (c) (d)

Q: What is the woman mainly doing in the conversation?

M: Shouldn't you be studying? Why are you watching TV?

W: I'm not – it's on for the background noise.

M: But don't you find it distracting?

W: Actually, it helps. I can't study in silence.

M: Still, won't you retain more without it on?

W: I've gotten all A's this semester studying like this.

Q: What is the woman mainly doing in the conversation?

(a) Justifying her study habit

(b) Complaining about being distracted by the TV

(c) Advising the man to watch less TV

(d) Describing how to get straight A's

10. (a) (b) (c) (d)

Q: What is the conversation mainly about?

W: Is that jacket leather?

M: Nah, it's fake. I don't wear leather.

W: Oh, since it's expensive?

M: No, I avoid it for moral reasons.

W: Really? But it's so fashionable.

M: Well, I think it supports animal cruelty.

Q: What is the conversation mainly about?

(a) Why leather jackets are fashionable

(b) The high price of leather jackets

(c) Why real leather is better than fake leather

(d) The man's choice not to wear real leather

11. (a) (b) (c) (d)

Q: What can be inferred about the man from the conversation?

W: Would you recommend laser eye surgery?

M: Absolutely. Are you thinking about doing it?

W: Maybe. My contact lenses have been irritating my eyes.

M: I had the same problem, and I didn't want to wear glasses.

W: I hate glasses, too. It'd be great to not need anything in order to see.

M: Well, everyone has different results, but I've never looked back.

Q: What can be inferred about the man from the conversation?

(a) He performs laser eye surgeries.

(b) He has had laser eye surgery.

(c) His contact lenses were too expensive to maintain.

(d) His eyesight is worse than the woman's.

12. (a) (b) (c) (d)

Q: What can be inferred from the conversation?

W: How were the football tryouts?

M: The pickings were slim. It ended a lot earlier than I expected.

W: Not much athletic talent among the students?

M: I wouldn't know. They're not trying out.

W: Maybe they're discouraged by last year's state championship.

M: That was a fluke! They should be begging to join a team with a record as good as ours.

Q: What can be inferred from the conversation?

(a) The man is trying to recruit coaches for the football team.

(b) The team made a good showing at last year's state championship.

(c) The woman assisted the man with the football tryouts.

(d) Fewer people tried out for the team than the man expected.

2-2 Written Mode

13.

M: Oh, this sandwich has cheese on it.

W: You asked for no cheese, right?

M: Yeah. Should I send it back?

W: I would if I were you.

M: I feel like I'm being fussy, though.

W: Just politely point out the mistake to the waiter.

Q: What is the woman's main advice to the man?

(a) To be more polite to the waiter

(b) To complain about the taste of the food

(c) To take more time to enjoy his meal

(d) To have the restaurant fix his order

14.

M: Welcome to my new apartment.
W: It's smaller than your old place but looks cozy.
M: It's all I could afford since rents are skyrocketing.
W: Yeah, a lot of people are downsizing.
M: Even this was barely in my price range.
W: And your old place was twice the size!

Q: What are the man and woman mainly discussing?

(a) Whether a bigger apartment is worth the price
(b) How much it would cost for the man to move
(c) Whether the man will have to move again
(d) How rent prices forced the man to rent a smaller apartment

15.

M: We've narrowed our list of job candidates to three.
W: Great. Should I set up their flights and interviews?
M: Yes. And please prepare summaries for the hiring committee.
W: Sure. And do you need anything else?
M: Please attach their resumes and cover letters.
W: I'll do that right away.

Q: What can be inferred from the conversation?

(a) The woman is the head of the hiring committee.
(b) The company plans to hire more than three new employees.

(c) The job applicants will be travelling to attend the interviews.

(d) The hiring committee has already interviewed some candidates.

16.

M: You know that liquidation sale I was telling you about?
W: Yeah, I'm going there tomorrow to buy a camera.
M: Well, it actually ended yesterday.
W: Oh, I was really looking forward to it.
M: I'm sorry. I must have misread the ad.
W: That's OK. Maybe I'll look for a camera online.

Q: What can be inferred from the conversation?

(a) The woman is buying a camera for the man.

(b) The sale will happen again at a future date.

(c) The man gave the woman incorrect dates for the sale.

(d) The woman has found a camera sale online.

Form B

<Dialogue-completion>

1-1 Written Mode

1.

M: Are you ordering the hamburger or burrito?
W: I don't know; they both look good.
M: How about we share?
W: _____

(a) Good thinking. We can split one of each.

(b) I don't know what they serve.

(c) No, the meal is on me.

(d) Sure, we can get something else.

2.

M: Did you get everything on the grocery list?
W: Almost – they were out of cherries.
M: But I need those for the muffins.
W: _____

(a) OK, I'll take them out.

(b) Just don't include them.

(c) They didn't have those, either.

(d) No, that recipe needs cherries.

3.

M: I’m considering hiring a cleaning service.
W: Aren’t they expensive?
M: Yeah, but I’m too busy to keep my house tidy.
W: _____

(a) That’s much longer than I spend.

(b) Then maybe it’s worth the price.

(c) I usually keep mine at home.

(d) Ask for your money back.

4.

M: Professor, can I turn in my essay tomorrow?
W: You know it’s due today, right?
M: Sorry, I’ve been ill all week.
W: _____

(a) I’ll give it to you later.

(b) Sure, today works for me.

(c) Then I’ll make an exception just this once.

(d) I can’t return it today, anyway.

1-2 Aural Mode

5. (a) (b) (c) (d)

W: How's your job searching going?
M: I've submitted several applications.
W: But no offers yet?
M: _____
(a) No, but I'm staying hopeful.
(b) I'll decide soon enough.
(c) Right. I don't know which to choose.
(d) I haven't applied there yet.

6. (a) (b) (c) (d)

W: Can you come to Friday's basketball game?
M: Hmm… Friday's not great for me.
W: Don't tell me you've already got plans!
M: _____
(a) I do, but I'll try to change them.
(b) Even so, I hope you'll come.
(c) I can make it if you move it to Friday.
(d) Honestly, I won't forget.

7. (a) (b) (c) (d)

W: Hi, I placed an online order but accidentally gave the wrong address.

M: OK, do you know if it's been shipped yet?

W: I doubt it. I placed it a minute ago.

M: _____

(a) You can just order it online.

(b) We apologize for the mistake.

(c) Then I can help you change it online.

(d) Sorry, the order hasn't been shipped.

8. (a) (b) (c) (d)

M: Can you give this file to Aaron when he gets back on Monday?

W: Sure. He's not here today?

M: He's on vacation. You didn't know he was gone?

W: _____

(a) No, he already asked me about it.

(b) Still, I'll be glad to have him back.

(c) It has been a while since I told you.

(d) I've been so busy that I didn't notice.

<Q&A>

2-1 Written Mode

9.

M: Shouldn't you be studying? Why are you watching TV?
W: I'm not – it's on for the background noise.
M: But don't you find it distracting?
W: Actually, it helps. I can't study in silence.
M: Still, won't you retain more without it on?
W: I've gotten all A's this semester studying like this.

Q: What is the woman mainly doing in the conversation?

(a) Justifying her study habit

(b) Complaining about being distracted by the TV

(c) Advising the man to watch less TV

(d) Describing how to get straight A's

10.

W: Is that jacket leather?
M: Nah, it's fake. I don't wear leather.
W: Oh, since it's expensive?
M: No, I avoid it for moral reasons.
W: Really? But it's so fashionable.
M: Well, I think it supports animal cruelty.

Q: What is the conversation mainly about?

(a) Why leather jackets are fashionable

(b) The high price of leather jackets

(c) Why real leather is better than fake leather

(d) The man's choice not to wear real leather

11.

W: Would you recommend laser eye surgery?
M: Absolutely. Are you thinking about doing it?
W: Maybe. My contact lenses have been irritating my eyes.
M: I had the same problem, and I didn't want to wear glasses.
W: I hate glasses, too. It'd be great to not need anything in order to see.
M: Well, everyone has different results, but I've never looked back.

Q: What can be inferred about the man from the conversation?

(a) He performs laser eye surgeries.

(b) He has had laser eye surgery.

(c) His contact lenses were too expensive to maintain.

(d) His eyesight is worse than the woman's.

12.

W: How were the football tryouts?
M: The pickings were slim. It ended a lot earlier than I expected.

W: Not much athletic talent among the students?
M: I wouldn't know. They're not trying out.
W: Maybe they're discouraged by last year's state championship.
M: That was a fluke! They should be begging to join a team with a record as good as ours.

Q: What can be inferred from the conversation?

(a) The man is trying to recruit coaches for the football team.
(b) The team made a good showing at last year's state championship.
(c) The woman assisted the man with the football tryouts.
(d) Fewer people tried out for the team than the man expected.

2-2 Aural Mode

13. (a) (b) (c) (d)

Q: What is the woman's main advice to the man?
M: Oh, this sandwich has cheese on it.
W: You asked for no cheese, right?
M: Yeah. Should I send it back?
W: I would if I were you.
M: I feel like I'm being fussy, though.
W: Just politely point out the mistake to the waiter.
Q: What is the woman's main advice to the man?
(a) To be more polite to the waiter
(b) To complain about the taste of the food

(c) To take more time to enjoy his meal

(d) To have the restaurant fix his order

14. (a) (b) (c) (d)

Q: What are the man and woman mainly discussing?

M: Welcome to my new apartment.

W: It's smaller than your old place but looks cozy.

M: It's all I could afford since rents are skyrocketing.

W: Yeah, a lot of people are downsizing.

M: Even this was barely in my price range.

W: And your old place was twice the size!

Q: What are the man and woman mainly discussing?

(a) Whether a bigger apartment is worth the price

(b) How much it would cost for the man to move

(c) Whether the man will have to move again

(d) How rent prices forced the man to rent a smaller apartment

15. (a) (b) (c) (d)

Q: What can be inferred from the conversation?

M: We've narrowed our list of job candidates to three.

W: Great. Should I set up their flights and interviews?

M: Yes. And please prepare summaries for the hiring committee.

W: Sure. And do you need anything else?

M: Please attach their resumes and cover letters.

W: I'll do that right away.

Q: What can be inferred from the conversation?

(a) The woman is the head of the hiring committee.

(b) The company plans to hire more than three new employees.

(c) The job applicants will be travelling to attend the interviews.

(d) The hiring committee has already interviewed some candidates.

16. (a) (b) (c) (d)

Q: What can be inferred from the conversation?

M: You know that liquidation sale I was telling you about?

W: Yeah, I'm going there tomorrow to buy a camera.

M: Well, it actually ended yesterday.

W: Oh, I was really looking forward to it.

M: I'm sorry. I must have misread the ad.

W: That's OK. Maybe I'll look for a camera online.

Q: What can be inferred from the conversation?

(a) The woman is buying a camera for the man.

(b) The sale will happen again at a future date.

(c) The man gave the woman incorrect dates for the sale.

(d) The woman has found a camera sale online.

Appendix B. Survey

시험 후 설문지_ Form A

참가자 정보

이름: _____ 전공: _____ 학년: _____

공인 영어시험 점수: 텝스 _____; 토익 _____; 토플 _____

- **다음 질문을 읽고 해당되는 번호(1에서 5 중 하나) 또는 항목에 동그라미 쳐주세요.**

1. 각 문항 유형이 얼마나 쉽거나 어렵게 느껴졌나요?	
파트 1-1 (대화 완성하기 - 선택지 4개 듣고 답하기)	매우 쉽다 1 2 3 4 5 매우 어렵다
파트 1-2 (대화 완성하기 - 선택지 4개 읽고 답하기)	매우 쉽다 1 2 3 4 5 매우 어렵다
파트 2-1 (질문에 적절한 답 고르기 - 선택지 4개 듣고 답하기)	매우 쉽다 1 2 3 4 5 매우 어렵다
파트 2-2 (질문에 적절한 답 고르기 - 선택지 4개 읽고 답하기)	매우 쉽다 1 2 3 4 5 매우 어렵다
2. 각 문항 유형이 청해 능력을 얼마나 잘 평가한다고 생각하나요?	
파트 1-1 (대화 완성하기 - 선택지 4개 듣고 답하기)	전혀 평가하지 못함 1 2 3 4 5 매우 잘 평가함
파트 1-2 (대화 완성하기 - 선택지 4개 읽고 답하기)	전혀 평가하지 못함 1 2 3 4 5 매우 잘 평가함
파트 2-1 (질문에 적절한 답 고르기 - 선택지 4개 듣고 답하기)	전혀 평가하지 못함 1 2 3 4 5 매우 잘 평가함
파트 2-2 (질문에 적절한 답 고르기 - 선택지 4개 읽고 답하기)	전혀 평가하지 못함 1 2 3 4 5 매우 잘 평가함
3-1. 파트 1(대화 완성하기)에서 선택지를 읽고 답하는 방법과 선택지를 듣고 답하는 방법 중 어떤 유형을 더 선호하나요?	(선택지 시험지에 제시함- 읽고 답하기) / (선택지 시험지에 제시하지 않음 - 듣고 답하기)
3-2. 그 이유는 무엇인가요?	
4-1. 파트 2(질문에 적절한 답 고르기)에서 질문과 선택지를 시험지에 제시하는 방법과 시험지에 제시하지 않고 들려주는 방법 중 어떤 유형을 더 선호하나요?	(선택지 시험지에 제시함- 읽고 답하기) / (선택지 시험지에 제시하지 않음 - 듣고 답하기)

4-2. 그 이유는 무엇인가요?	
5. 다음 문장을 읽고 각 문장에 얼마나 동의하는지 표시해주세요.	
1) 나는 문제를 들려주는 방법이 쉬웠다. 문제와 선택지를 읽고 이해하지 못할 것을 걱정할 필요가 없기 때문이다.	전혀 동의하지 못함 1 2 3 4 5 매우 동의함
2) 나는 문제를 보여주는 방법이 쉬웠다. 문제와 선택지를 듣고 이해하지 못할 것을 걱정할 필요가 없기 때문이다.	전혀 동의하지 못함 1 2 3 4 5 매우 동의함
3) 나는 문제를 들려주는 방법이 어려웠다. 지문이 나오기 전에 선택지를 읽을 수 없었기 때문이다.	전혀 동의하지 못함 1 2 3 4 5 매우 동의함
4) 나는 문제를 보여주는 방법이 어려웠다. 선택지를 시간 안에 (빨리) 읽지 못했기 때문이다.	전혀 동의하지 못함 1 2 3 4 5 매우 동의함
5) 나는 들려주는 방식의 문제를 풀 때 선택지를 기억하는 데 어려움이 없었다.	전혀 동의하지 못함 1 2 3 4 5 매우 동의함
6) 보여주는 방식은 어려웠다. 왜냐면 읽기와 듣기를 동시에 해야 했기 때문이다.	전혀 동의하지 못함 1 2 3 4 5 매우 동의함
7) 들려주는 방식의 문제를 풀 때, 나는 모든 선택지를 들려줄 때까지 기다릴 필요가 없었다. 정답이 나오면 바로 고를 수 있었다.	전혀 동의하지 못함 1 2 3 4 5 매우 동의함

연구에 참여해 주셔서 진심으로 감사합니다!

국 문 초 록

본 연구는 문항 제시 방법 및 문항 유형이 한국인 대학생 영어 학습자들의 청해 시험 수행과 그에 대한 인식에 미치는 영향을 분석하고자 하였다. 영어 능숙도에 따라 세 그룹으로 나누어 모집한 115명의 한국인 대학생들이 본 연구에 참여하였으며, 모든 참여자들은 문항 제시 방법(보여주기와 들려주기)과 문항 유형(담화완성과 질의응답)에 따라 총 16개의 문항으로 구성된 청해 시험에 응시하였다. 청해 시험이 모두 끝난 후에 각 형식에 대한 인식을 묻는 설문지 작성이 이루어졌으며, 각 그룹의 약 10%인 총 12명의 참여자들이 반추하기 기법(stimulated recall)을 사용한 인터뷰에 응하여 각자의 문제해결 과정을 제공하였다.

연구 결과, 하 그룹의 경우 문항 유형에 따른 점수의 차이는 보이지는 않았지만, 문항 제시 방법에 있어서 들려주기 방법보다 보여주기 방법에서 더 높은 점수를 획득했다. 이와 관련하여 하 그룹은 들려주기 방법이 보여주기 방법보다 훨씬 어렵다고 느꼈으며, 보여주기 방법을 선호하였다. 설문지와 반추하기 기법을 사용한 인터뷰를 통해 능숙도가 낮은 집단이 들려주기 방법을 어려워하는 이유를 알아보았고, 이는 들려주기 방법이 고도의 집중력과 좋은 기억력을 요하기 때문임을 확인했다. 그러나 보여주기 방법에서도 선택지를 시간 안에 읽을 수 있는 읽기 능

력을 요하는 등 듣기 능력의 구인에 적절하지 않은 요소가 드러남을 확인할 수 있었다.

한편, 중 그룹과 상 그룹의 경우 문항 유형 중 질의응답 유형에서보다 담화완성 유형에서 더 높은 점수를 받았지만, 문항 제시 방법에 따른 유의미한 차이는 나타나지 않았다. 그러나 두 집단 역시 들려주기 방법이 보여주기 방법보다 훨씬 어렵다고 응답하였으며, 보여주기 방법에 대해 하 집단보다도 더 강한 선호도를 보였다. 설문지와 반추하기 기법을 사용한 인터뷰 결과, 능숙도가 높은 집단은 보여주기와 들려주기 모두에서 비슷하게 잘 할 수 있음에도 불구하고 질문과 선택지를 듣는 것보다 읽는 것에 자신감을 가지고 있었고, 미리 다음 문제의 선택지를 읽고 듣기 내용을 예상할 만큼 충분히 빠른 속도로 읽는 것이 가능했으며, 이러한 방법을 자주 사용하였다.

결과를 종합해보면, 청해 능력 평가에 있어서 구인에 적절하지 않은 요소(construct-irrelevant factor)로 가장 주요한 것은 들려주기 방법에서는 기억력, 보여주기 방법에서는 읽기 능력이었다. 청해 능력 평가의 결과에 기억력과 읽기 능력이 큰 영향력을 미치는 것은 타당하다고 볼 수 없으므로, 평가를 개발할 때 이러한 요소들의 영향을 최소화하는 것이 중요하다. 이러한 결과에 근거하여 본 연구는 선다형 듣기 평가 개발과 선정에 있어서의 시사점과 이후 연구에 대한 제언을 결론부에 제

시한다.

주요어: 선다형 문항, 제2언어 듣기, 문항 제시 방법, 문항 유형

학 번: 2014-22961